STERN, LUCAS ERIC, Ph.D. A Psychometric Evaluation of the Student Teacher Assessment System at UNCG Using Structural Equation Modeling. (2016)
Directed by Dr. Terry Ackerman. 155 pp

Student teachers are evaluated based on a number of criteria at the University of North Carolina at Greensboro (UNCG). Among these criteria are the evidence portfolios, the Candidate Disposition Assessment Process (CDAP), and the Teacher Growth and Assessment for Pre-Service Profile (TGAP) instruments. These instruments attempt to measure teacher skills and attitudes at various points throughout a student teacher's progress leading up to graduation. Structural equation modeling was used first to compare the appropriateness of five confirmatory factor analysis models when the data is fit to each of them. Next, the most appropriate model was used to explore the quality of the items. Differences were examined between teacher candidate evaluators in cases where multiple raters exist. Finally, a differential item functioning (DIF) analysis will compare the way items were interpreted by the evaluators of elementary teacher candidates and middle secondary teacher candidates by the use of a multiple indicator-multiple cause model.

The six-factor correlated model best represented the data from the assessment system of teacher candidates at the UNCG. Details of the model showed good evidence for the reliability of all six factors, however further data needs to be collected and preserved in order to draw conclusions about the instruments capacity to distinguish accurately teacher candidates whose abilities in at least one of the six areas being measured would put them near the cut score. Evidence was found of high inter-rater reliability between the supervising teacher and the on-site teacher evaluator for scores on

the CDAP instrument, however differences were found between these raters for the TGAP instrument. In the demonstration of DIF detection, a few items were flagged as potentially being interpreted as significant differences in ratings between secondary/middle evaluators and elementary evaluators.

A PSYCHOMETRIC EVALUATION OF THE STUDENT TEACHER ASSESSMENT SYSTEM AT UNCG USING STRUCTURAL EQUATION MODELING

by

Lucas Eric Stern

A Dissertation Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2016

Approved by

_____
Committee Chair

DEDICATION

I dedicate this dissertation to Jessica my amazing wife who supported me in every way along this journey and to Larry my father whose help and support in this project and in life continues to demonstrate both love and loyalty.

APPROVAL PAGE

This dissertation written by Lucas Eric Stern has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair _____

Committee Members _____

_____

_____

_____
Date of Acceptance by Committee

_____
Date of Final Oral Examination

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

Introduction

Evaluation of student teachers at the conclusion of their program at the University of North Carolina at Greensboro (UNCG) at the School of Education involves the use of several instruments. The assessments that were used in this study are the evidence portfolios, the Candidate Disposition Assessment Process (CDAP), and the Teacher Growth and Assessment for Pre-Service Profile (TGAP). The evidence portfolios are measured once during the final evaluation of a student teacher candidate and include a total of 51 items which measure four constructs. These constructs include depth of content, pedagogical knowledge and skills with a focus on planning, impact on student learning, and leadership advocacy and professional practice. Each of these items is scored on a three-point scale. The CDAP is used three times throughout the course of a teacher candidates' training, and measures certain aspects of the candidate such as ethical behavior, receptiveness to feedback, collaboration, and responsibility. Both the CDAP and TGAP assessments are measured on a six-point scale with averages of multiple evaluators allowed. The TGAP is also used three times throughout the teacher candidates' experience, and measures candidate growth in planning, instruction, assessment, and student motivation and management using 18 items.

The following research presents an evaluation of this assessment system by using structural equation modeling to: 1) find the most appropriate model to represent these data; 2) address the quality of the items in terms of their ability to measure their intended construct(s); 3) explore the agreement between multiple evaluators and to detect the potential presence of differential item functioning (DIF) or measurement bias when interpreted by evaluators. Results of this research will provide evidence, examples, and guidelines for the application of statistical methods, specifically structural equation modeling, which could be used to evaluate aspects of a student teacher assessment system. This research will demonstrate an application of structural equation modeling to serve both the UNCG and other institutions throughout the country, that seek the improvement and confirmation of the quality of their own teacher evaluation assessment systems.

Statement of the Problem

Licensure of a student teacher candidate requires a successful student teaching experience as well as the completion of various requirements or assessments. In the past, the role of teacher education faculty in providing summative assessments of their teacher candidates has not been well respected, even by the faculty (Farkas, Johnson, & Duffett, 1997; Raths & Lyman, 2003). Often factors that have led to this lack of credibility include questionable measurement instruments and untrained evaluators (Barrett, 1986). Another problem with the accurate evaluation of student teachers is rooted in the fact that negative formative evaluations are often very difficult for an evaluator to record. Raths and Lyman (2003) made the argument that, in many cases, the university supervisor acts

as both coach and formative evaluator. Near the beginning of a semester, the university supervisor will offer guidance and support trying to help with a student's struggles and weaknesses, and then, as the semester comes to a close, a struggling student may have those same weaknesses that were shared early in the semester used in the decision not to recommend the student for licensure. The university supervisor as the evaluator is then faced with either passing the student anyway or having to answer tough questions from the student: "Why didn't you help me with these issues when we knew about them earlier?" However, the positive influence of supervising teachers tends to exceed that of other influences such as cooperating teacher or principals at the site of their student teaching experience (Farrell, 2008). In some cases, students feel it is difficult to get timely feedback from supervisors, or they are too focused on pleasing them and passing the practicum to learn the skills that they need to be successful in the profession (Ochieng'Ong'ondo & Borg, 2011; SelormSosu, Paddy, AsantewaaMintah-Adade, & Ativui, 2014). In a study by Borko and Mayfield (1995), it was found that most cooperating teachers and university supervisors played a minor role in student teachers' growth in teaching. Another study by Merç (2015) examined whether or not student teachers were satisfied with the methods used to evaluate their performance. After analyzing his findings Merç (2015) suggested that cooperating teachers in general need better professional qualifications to effectively assess student teachers. He argued that their ability to reliably and accurately measure student performance could be called into question. It is important for university supervisors to internalize the evaluation criteria so that they are consistent in their appraisal of the student teacher's performance. Also, it is

important that too much emphasize is not placed on written reports but multiple techniques are used such as observation so that a more complete picture of the student teachers' experiences can be collected and used for feedback and assessment.

Statement of the Purpose

This research demonstrates how structural equation modeling can be applied in the support of the continuous improvement of student teacher preparation and assessment systems. While often based on a similar set of standards, assessment systems vary depending on the university where the license is earned. This research involves examining a specific collection of assessment tools used for licensure recommendation decisions at the school of education at the UNCG. The use of responsibly and correctly applied research tools such as structural equation modeling (SEM) can provide support and feedback to a variety of assessment systems.

Research Questions

The research questions driving this evaluation begin with one question: What confirmatory factor analysis (CFA) model can be used to appropriately represent the structure of this assessment system given the data? This assessment system is used to evaluate teacher candidates as a whole and potentially measures a latent ability trait that will be designated teaching capacity. Each of the evidence portfolios, the CDAP, and the TGAP instruments measure what is believed to be related, yet different aspects that make up the larger ability. Preliminary analysis revealed that the data exhibit characteristics of both a unidimensional data set and a multidimensional data set. The data are expected to be represented best by a bi-factor model that allows for the display of each item's

contribution to both a general factor (teaching capacity) and a specific factor such as the TGAP or the CDAP. In order to support the use of this model, five confirmatory factor analyses were performed and the goodness of fit compared between models. The five models included a single factor model where each item was only allowed to be influenced by the general factor, a bi-factor analysis where each item was allowed to be influenced by one general factor and the remaining variance explained by one specific factor, a six-factor model which modeled each items' contribution to the specific factor to which the item was assigned if the factors TGAP, CDAP, and the evidences were independent with regard to each other, a fourth model with the same six factors where correlations were estimated between the factors, and, finally, a higher order model where the specific factors loaded directly onto the general factor and the relationship between them was explained primarily by the general factor. Several indices of fit were considered and compared with parsimony in mind. The results of this analysis should give us some insight as to the most reasonable factor structure by which to represent this data.

The second research question considered was the following: When a structure was imposed on the model what do the results reveal about the items and the assessment? Parameters were interpreted from the CFA model that exhibited the best fit to the data including the standardized loadings and the threshold parameters for each item. The loadings provided some insight into how well each item contributes to a specific factor and can be used as a measure related to item quality. The threshold parameters will give us an idea of the difficulty of each item as the threshold represent for multiple categories the level of the latent trait required to transition from one category to the next.

The third question that is of interest to this study asks the following: Do differences exist between the scores given by the supervising teacher and the cooperating teacher and, if so, what is the nature of this difference? This question was addressed using a multi-group structural equation model using only data where multiple evaluator scores were present. This test looked at invariance on multiple levels to see if the CFA model held across groups. Invariance was tested across the loadings, thresholds and latent means as appropriate across groups. Data only existed for multiple evaluators in the TGAP and CDAP instruments and only from certain programs so the number of candidates included were less than the number in the first model. Also, since only two potential factors were included, this model was not the same as the model for the full data set but will correspond to that model if only two factors were present.

The last research question that this study addressed, discovered whether differences exist in the way items are interpreted and scored depending on whether the teacher candidate is in an elementary education or special education program or in a middle grade/secondary program. This division was intended to serve as an example of the type of group division that could be of interest to teacher candidate evaluation programs. When group membership is determined, it is important to consider sample size when determining the appropriateness of the methodology. This question will be answered by using structural equation modeling to perform a DIF analysis. Specifically, a multiple indicator multiple cause (MIMIC) model will be used to test for DIF. Evaluator scores from programs considered to be elementary education will make up one group and evaluator scores that are considered to be in middle/secondary education will

make up the other group. These evaluator types will form the reference and focal groups respectively. This analysis, rather than being another multi-group analysis, will have an indicator variable that represents group membership. It will also have all latent variables load onto this indicator variable to partial out this effect, and then any loadings that are significant onto the indicator variable from the individual items will be considered evidence for DIF for that item. The results of this analysis revealed whether some items seem to be interpreted differentially by the two groups. These results will also be compared to another method of DIF testing using Polysibtest (Walker, 2001). These analyses will demonstrate applications of structural equation modeling to teacher assessment as well as provide specific insight into the assessment system at UNCG.

The Assessments

The assessments that with be used in this study include the evidence portfolios, the CDAP, and the TGAP. The evidence portfolios are made up of six evidences, each representing a project that a student completes and a supervisor or combination of supervisors will collect and assess based on a rubric. Portfolios provide a unique contribution to the assessment of student teacher candidates beyond that of student teacher grades or Praxis results (Simpson, 2004). Because the assessment of teacher dispositions is a requirement for the National Council for the Accreditation of Teacher Education (NCATE) institutions nationwide are working to clearly define and assessment teacher candidate dispositions (Almerico, Johnston, Henriott, & Shapiro, 2011). Kim, Micek, and Grigsby (2013) addressed the difficulty in teaching attitudes to teacher candidates and emphasized the importance of modeling the desired behavior and

attitudes. Evidence One and Evidence Four will be excluded from this analysis and only evidence two, three, five, and six will be included.

There were two evidence portfolios used in the assessment system but not in this analysis. Evidence One is called breadth of knowledge, and is demonstrated by students producing proof of their successful completion of the Praxis II with a passing score as well as a transcript demonstrating the completion of specific coursework required for initial teaching licensure. Since the teacher candidates represented by this data set are all completers, the expected variability of scores for this evidence will be zero. Evidence Four represents the successful completion of the student teaching experience. This evidence is also graded as pass/fail and, like in evidence 1, the variability is expected to be near zero. Therefore, since the only story the scores from these evidences should tell us is that all completers passed this criterion, further analysis was not pursued.

Evidences Two, Three, Five, and Six are each made up of multiple criterions, each scored on a three-point scale. A score of 1 means the candidate did not meet the requirements for that item, a score of 2 represents proficiency, and a score of 3 represents the candidate exceeding expectations. The same basic scale is used for items on a six - point scale where 1-2 represents a failure to meet the requirements, a 3-4 represents proficient, and a 5-6 exceeds expectations. Evidence two, which is the in-depth inquiry project, requires the student to complete an approved project appropriate for their field, which is then assessed based on performance using eight criterions. This demonstrate the student's competency in content, depth, rigor, and presentation. Evidence Three is completed through the successful crafting and implementation of both a unit and daily

lesson planning and is made up of 15 criteria. Evidence Five measures the impact on student learning and is demonstrated through 20 criteria involving planning, instructional monitoring, teacher adaptation, and data collection and analysis. Evidence Six is leadership advocacy and professional practice, which includes the completion of a project evaluation based on eight criteria. The CDAP assessment measures candidate dispositions at least three times throughout the training experience: once at the beginning of their training used as a baseline, again before their student teaching, and finally at the end of their student teaching. Only the final assessment is considered summative but earlier assessments help to highlight areas a teacher candidate might be struggling with in time for productive intervention and feedback. The TGAP is also used at least three times during the course of teacher preparation and measures 18 criteria including growth, and progress in the areas of planning, instruction, assessment, and student motivation and management. Like the disposition assessment, only the final assessment was considered summative. The six evidences are aligned with the North Carolina Professional Teaching Standards (NCPTS, North Carolina Professional Teaching Standards, 2011), and the TGAP is aligned with the Interstate Teacher Assessment and Support Consortium (InTASC) (Council of Chief State School Officers, 2015) standards (Assessment & Support Consortium, 2011). A detailed description of the assessment criteria and their alignment with state and national standards is provided in Figure 1 in Appendix A.

Theoretical Framework

In the past there has been reason to question the quality of teacher assessment (Barrett, 1986; Farkas et al., 1997; Merç, 2015; Raths & Lyman, 2003). Some methods

have sought to improve teacher candidate assessment quality through better definition of assessment concepts (Almerico et al., 2011), examining effects of evaluator training (McIntyre & Killian, 1987), and efforts to analyze candidate evaluation instruments (Benjamin, 2002; Danielson, 2011; Pecheone & Chung, 2006; Voss, Kunter, & Baumert 2011). Danielson (2013) established a framework of teaching that emphasizes the four major categories of planning and preparation, the classroom environment, and instruction and professional responsibilities. This framework has been used to develop rubrics for the evaluation of multiple aspects of these categories. This research seeks to support the practice of continuous improvement in education as well as demonstrate the use of advanced statistical methods to support student teacher assessment instruments in licensure recommendation decisions. Historically, local decisions on teacher certification requirements have moved toward certification decisions, becoming dependent on the fulfillment of requirements at a state and national level. This paper begins with a brief review of the history of the profession and how teacher certification and the accreditation of teacher preparation programs have developed over the years. The next section explores the relationship of the teacher candidate certification process to the educational experience of the students taught by the certified teacher. Following this, a summary of similar research will be provided as well as a discussion on how this specific research fits in the context of other research with similar intent and into the larger picture of educational assessment. Finally, a description of methodologies used in this research are provided.

UNCG has been training educators for more than a century (UNCG, 2015). One of the key requirements today for the completion of a licensure program in teaching from UNCG includes a passing score on multiple instruments intended to measure student achievement, competence, teacher candidate dispositions, and growth. The results of this research will demonstrate how structural equation modeling can be applied to specific questions faced in the assessment of teacher candidates, which can be applicable to other educational institutions based on their needs.

Overview of Methods

Preliminary analyses explored the reliability of the assessment and the factors, the dissattenuated correlation between each factor, and checked for normality with descriptive statistics including P-P plots, skewness, and kurtosis. To test for reliability Cronbach's alpha was calculated in SPSS for each construct and the all constructs combined into one assessment. Coefficient alpha is generally thought of as having a range from zero to one and a higher value represents a better reliability (Cronbach, 1951). The dissattenuated correlation coefficients were calculated by dividing the correlation between each variable by the geometric mean of the reliability coefficients of each assessment using this formula:

$$R_{xy} = r_{xy} / \sqrt{r_{xx} * r_{yy}} \qquad (1)$$

In the equation, r_{xy} is the correlation coefficient and r_{xx} and r_{yy} represent the reliabilities of the assessments (Muchinsky, 1996; Spearman, 1904). This formula uses the reliabilities to remove the measurement error due to unreliability and provides an

estimate of the correlation if both assessments were perfectly reliable. Calculations for reliability, skewness, kurtosis, and P-P plots were performed in the Statistical Package for the Social Sciences (SPSS) (IBM Corp., 2012).

The preliminary analysis revealed that the reliability is very high (>0.9) for both the combined assessment and each of the six constructs (the four evidence portfolios, the CDAP, and the TGAP). Table 1 shows the correlation coefficients corrected for attenuation of factors on the upper triangle, Pearson correlations between factors on the lower portion and reliabilities on the diagonal. One note about the reliabilities for the evidences is that often times reliability is higher when an instrument has more items as a function of the formula used. In this case, evidence two and evidence six have eight items each while evidence three and five have 15 items and 20 items respectively. As is not surprising in this case, evidence Three and Five demonstrate the higher reliabilities as compared to evidence Two and Six, which contain fewer items. Evidences are moderately correlated with other evidence factors have low to moderate correlation with two exceptions. The TGAP and CDAP seem to be highly correlated and evidence 2 seem to have almost no correlation with either of the TGAP or CDAP factors. Whereas the TGAP and CDAP factors are highly correlated, they are theoretically intended to measure different things. The TGAP measures teacher performance, while CDAP measure dispositions. For this reason, these factors are both retained and analyzed separately rather than combined into one factor.

Table 1

The Correlation Coefficients Corrected for Attenuation of Factors

	Ev2	Ev3	Ev5	Ev6	CDAP	TGAP
Ev2	.93	.37	.47	.41	.08	.06
Ev3	.35	.97	.43	.38	.22	.21
Ev5	.45	.42	.97	.49	.13	.15
Ev6	.38	.36	.46	.93	.17	.20
CDAP	.07	.21	.13	.16	.95	.77
TGAP	.05	.21	.14	.19	.74	.98

The TGAP is divided into four sub-constructs including planning, instruction, assessment and student motivation and management. While the correlations of items within the TGAP were high, correlations of variables within the sub constructs of the TGAP did not stand out as being consistently higher than those paired across sub-constructs. Therefore, the sub-constructs of the TGAP will not be explored further in this analysis and will be combined into the single construct. The test for normality revealed that based on the skewness, kurtosis, and P-P plots, univariate normality is violated in several places. The data exhibited characteristics of multi-dimensionality based on the correlations corrected for attenuation and on the results of a principal components analysis.

The first analysis compared the fit indices of five different structural equation models including a single factor model, a bi-factor model (Holzinger & Swineford, 1937), a six-factor model orthogonal model, a higher order model, and a six-factor correlated model, similar to the comparison done by Yang et al. (2013) in the modeling of acute stress response. Once the factor structure was confirmed, the loadings and the

thresholds can be used to draw conclusions about the items and the assessments. Assuming the one-factor model is the retained model, the factor analysis should produce similar conclusions to that of Samejima's (1969) item response theory (IRT) graded response model since the data are categorical in nature. Analysis of these parameters will provide insight into the performance of the items and the assessments.

Next, in order to compare scores given to students by multiple raters, a multi-group SEM model will be used. One group in this model is represented by the university supervising teacher and the other group by the on-site teacher evaluators. Invariance across groups will be tested for the general structure, the loadings, the thresholds and if invariance holds, the latent variable means. At least one evaluator graded each assessment for each teacher candidates. In the case of multiple evaluators, scores were averaged to produce the final result. Because only TGAP and CDAP scores had multiple evaluators and then only within certain programs, the model corresponded to the retained model from the previous analysis. The results provide insight into any difference between the two groups that may exist as well as where that difference exists. SEM analyses are tailored to large samples sizes and work best when these are available. Based on the research done about the appropriate minimal sample size for a CFA; Westland (2010) suggested that the best way to approach this question is to consider the ratio of observed variable to latent constructs. Another rule of thumb states that a ratio of number of participants to number of model parameters ideally should be about 20:1; a practical goal in practice might look more like 10:1 and ratios less than 5:1 may produce unstable estimates (Suhr, 2006). The final analysis performed in this research was a test

for DIF between elementary education evaluators and middle/secondary evaluators using an SEM modeling framework. Using a MIMIC model, this analysis identified items that can be interpreted differentially by the two groups. The results of these collective analyses were summarized and presented to address each research question.

Summary

The goals of this research include providing examples of how advanced statistical methods can be applied to address the issues and challenges faced when implementing a teacher candidate evaluation system as well as to provide psychometric support for the instruments involved in the certification process for new teachers at UNCG. Evidence of the quality of teacher candidates helps to serve and support good teachers, education programs and employers alike.

CHAPTER II

REVIEW OF THE LITERATURE

A Brief History of Student Teaching Evaluation

The art of passing down knowledge to future generations to enhance and preserve life is an ancient practice. Although some values have been held in high regard throughout the ages, other societal values match both the life styles and culture of the people in which they are taught. In the time of Confucius (551 BC – 479 BC), the Chinese instructor emphasized six arts including archery, calligraphy, computation, music, chariot driving, and ritual. In the early centuries, countries like Greece and Rome recognized the value of education and often would have teachers or educated slaves instruct their children from home, or pay for education in private schools. Some common curriculum focuses included gymnastics, music, and literacy. Development of skills such as rhetoric, mathematics, logic, and politics were primarily reserved for the wealthy. In the Greek city-state of Sparta education was limited almost exclusively to physical training and combat tactics. In many families of ancient society, a child would be taught a trade by their father and this was as close to a formal education as they received. Starting around the 9th and 10th century, some of the oldest universities in the world began to open their doors to students in Morocco and Egypt. Still others such as Oxford and Cambridge were founded in Europe centuries before America was established as a nation (History of Education, 2015).

A graduate of Cambridge and Puritan minister, John Cotton (1585-1652) played a large part in establishing the first public school in America, the Boston Latin School in 1635. This school taught Greek and Latin and educated 5 of the 56 signers of the Declaration of Independence. Harvard, which was founded a year later in 1636, was another of the first schools established in America. This school was better known for its emphasis on theology. Around this time, the only major requirement for becoming a teacher was that the candidate had attended school themselves and had a good reputation. Teaching was primarily a profession for men in the 1600s but that trend changed over time. During the Civil War a large shift occurred in the gender roles of the profession. Men returned from war to find women competently fulfilling their role, and doing so at a far lower salary. As pay was low for an educator before the war, facing potentially an even lower salary to reclaim their jobs caused many men to leave the profession. The moral standard for teachers was also very high: women were not allowed to marry and were forced to resign if they did. In 1929, 11 high school teachers were fired by the Kansas Board of Education for attending a local country club dance. Up until the 1800s, formal teacher training and tests of teacher quality in American were generally not emphasized in a young teacher's career. Teacher competency was generally established quickly if the candidate attended school and could pass as literate (The History of Education, 2015).

LaBlue (1960) considered two reasons for teacher certification. The first purpose is to continuously improve, as well as guarantee the best possible education for our students. This purpose is based on the assumption that the quality of a student's education is largely influenced by the skills and preparation of the educators, a topic that

will be addressed later in this paper. The second reason for teacher certification is to protect the reputation of both the teaching profession as well as the individual certified teacher from unqualified competition. In America before 1789, there were few laws that required public schools to be established and parents were generally free to determine independently how the education of their children would progress. In the South, a common practice was for parents to make their own private agreements with an instructor; conversely, in the North, a mother who decided to educate her child might make an agreement with several neighbors and teach a small group of children. This type of school was often referred to as a dame school. From the early colonial period until present day, LaBlue (1960) divided the history of teacher certification into four distinct periods. The first period covers early colonial times up until about 1789, a period that is described as bearing some concern for teacher certification but little concern for the qualifications of the individual candidates. A teacher was certified based on reputation or minimal educational experiences. The second period covers about 1789-1860. During this period, control of teacher certification began to shift to state authority and normal schools began to appear.

In 1823, Hall established the first normal school, which offered teachers a 2-year course of instruction in the art. On July 3, 1839 the state of Massachusetts funded the first state supported school designed exclusively for teacher preparation in Lexington (Harper, 1939). At the dedication of the school in 1846, Bates announced the following:

> provision for the education of the people of the state at the expense of the state was essential for progress and prosperity; that the people could be educated only in the common [public] schools; and lastly, that the common schools could have an adequate teaching force only if the education of their teachers were provided for by the state. (Harper, 1939, p. 10)

The third period of the history of teacher certification extends from about 1860-1910, covering a rise in normal schools, the establishment of teacher colleges, and the appearance of schools of education departments within liberal arts universities (LaBlue, 1960). It was during this period that a school was established in 1891 under the name of the North Carolina State Normal and Industrial School, with the primary purpose of training female educators under the leadership of Charles Duncan McIver as its first president. In 1919, the school was renamed the North Carolina College for Women (NCCW) and later in 1963 admitted its first male student, during which time period the name was changed again to the UNCG, the name by which it is known by today. Along with this new title, UNCG is known as having within the School of Education one of the oldest teacher education programs in the state (UNCG School of Education, 2015). The fourth period of the history of teacher certification covers the years since 1910 particularly since 1930 where major developments have occurred in the area of improving teacher certification standards (LaBlue, 1960).

Up until the 19th century, concern for teacher qualification often emphasized only moral character rather than subject knowledge or skill in the art of teaching. It was common for a potential teacher to be approved by a local minister in colonial times, which in part could be dependent on the candidate's religious beliefs and their similarity

to the priest's own views. With the common schools beginning to replace the charity schools in the 1830s and 1840s, two patterns of education emerged. The first was comprised of thousands of one-teacher schools serving smaller, rural districts while the second served more urban regions contained larger multi-classroom schools organized into school systems and controlled by elected or appointed boards of education (Angus, 2001). Many differences in the pay, and circumstances of these systems and teachers framed the context of how and by whom teachers should be trained and licensed to practice. Some of the first teacher certifications were awarded upon successful completion of an oral examination that was often administered by a member of the district board. At first these exams were short with the goal of establishing a basic degree of competency in the subject matter, but later state elected officials required longer, more detailed written exams that would grant the successful candidate a certificate to teach within the administering area for varying lengths of time. New York led the way in statewide certification in 1843 as it authorized the state superintendent to administer exams and issue certifications that were recognized throughout the state. Indiana and Pennsylvania followed in this practice about a decade later and by the end of the 19th century most states were certifying teachers at the state level. By the end of the 19th century 28 states accepted the completion of a normal school as sufficient for teacher certification while other states or counties required additional examinations. The information covered on teacher examination began with literary qualifications and expanded later to spelling, arithmetic, geography, history, and English grammar. In 1867, in Pennsylvania the addition of *professional knowledge,* or the practice of teaching,

was included to encourage study of the principles of the profession itself, as opposed to exclusively mastery of content knowledge. Over the last third of the century, standardization of these teacher qualifications took precedence over their further expansion. Certification became more and more centralized as town certification moved to county then to state, and certain questions on state exams became required which had previously been voluntary (Angus, 2001).

While it was largely agreed that education was an important government focus, government intervention in issues such as organization and financing were often met with resistance at a local level. With widespread belief that educating children was something that most people could do and furthermore that talent had more to do with innate gifting than training or a knowledge base, it was difficult for professional educators to influence change. Even the professional knowledge curriculum itself was often viewed as common sense. Some teaching philosophies and methods from Europe were implemented in America during this time that helped establish some structure to the developing profession. The monitorial system, developed by Joseph Lancaster in England in 1803, allowed a single teacher to educate several hundred students at a time through the use of older students assisting and acting as monitors in a strict curriculum. In1860, Edward Sheldon, the head of the normal school in Oswego, New York promoted object teaching, associated with the Swiss educator Johann Pestalozzi. Late in the 19th century Johann Hebart, a German educator, put forth a collection of principles dubbed the "new education" which later gave way to progressivism. Still none of these philosophies collected in the profession the unanimous support needed to establish them as scientific

foundations, which would have provided the credence needed in the eyes of the public to show that training in educational philosophy significantly improves teacher's skills (Angus, 2001).

Throughout the 19th century, the movement of certification authority and examination control transitioned from the town to county and county to the state. In New York, by 1888, the state superintendent was given the power to prepare the questions in examination regarding certification, and by 1894 was given authority to score the exams and establish cut scores. By 1899, New York because the first state to have a uniform system of teacher certification that was under state control. This trend soon spread to other states, but a large part of the slow progression of professional educators control over the rural education was the lack of reform policy proposed within country schools that did not insist on total reconstruction of the structure and governance of these schools. Reform suggestions often reflected a disliking of the one teacher schools of rural America, pushing only strategies involving the elimination of such schools via consolidation rather than simple reform with the basic structure still intact. Oftentimes the advocates of rural schooling were less cooperative when the only suggestion of reform required assimilation (Angus, 2001).

An expansion in the 20th century of the conversion of normal schools into colleges, combined with the recognition of education departments within the universities and educational degrees, marked a substantial turn in the respect for the training of teachers. With this respect came the opportunity of educational professionals to impact the American educational system in a meaningful way. The educational leadership,

consisting primarily of faculty in educational schools and superintendents, began to separate themselves from the classroom and began to tackle the larger issues of educational policy and the implementation of the scientific educational strategies and restructuring of education. These progressives targeted legislatures to affect state laws in furthering their objectives, and aided by their high degree of consensus were often successful. These administrators sought to eliminate local certification of teachers in an attempt to completely centralize control of teacher certification while pushing for longer more rigorous and often specialized training for educators. World War I led to a temporary shortage in teachers but ultimately led to several successful campaigns for increased teacher salary. World War II, however, caused one of the most dramatic shifts in the profession. Many left teaching to join the service, and many more emergency licenses were issued. By the war's end there had been issued roughly as many emergency licenses as the annual addition of teachers to the profession, which was just under 109,000 (Angus, 2001). Benjamin Frazier, the U.S. Office of Education's senior specialist in teacher education, was afraid that with the large amount of emergency certifications that had been issued during war time, the high standards that had taken so many years to fight for would be pushed back. The march for higher standards, however, resumed after the war without taking a major hit. From the period of 1940 – 1953, requirements for initial license rose in many states. A minimum of a 4-year college degree was now required in 25 state as opposed to only 9 in 1940. The standards for high school teachers was also raised as 40 states required 4 years of college and 5 states required 5 years. During this period more shifts occurred from local authority to state

authority in the realm of teacher certification, and by 1953 only Massachusetts, Illinois, and Missouri were left sharing power of certification with local officials. Most cities and colleges at this time that were granted certification power had higher requirements than the states. Only three states remained where examination was used without prerequisite training from teacher certification in the rural areas. In 1946, the National Education Association (NEA) created the National Commission on Teacher Education and Professional Standards (TEPS). This organization was created to return a voice to classroom teachers where they had often been suppressed by the college educational staff and other members of the educational trust over the past few decades. The first conferences held by this organization acknowledged two important aspects of teacher certification; firstly, to protect the public from incompetent teachers, and second to protect qualified teachers from unfair competition. TEPS soon gained representation in every state and pushed for both the minimum qualifications for teachers to be a bachelor's degree, as well as the elimination of certification exams. TEPS is also responsible for the creation of advisory councils to assist state certification officers, which, instead of representing mostly parties like deans of educational schools and normal schools, represented a much larger breath of stakeholders in the teaching profession. TEPS began the approved training approach where state departments would approve teacher-training institutions. Soon to follow was the creation of NCATE. In 1950, at a regional conference, Ralph MacDonald claimed that with the exception of a very few states, teacher training in the United States was a "travesty on professional education" (Angus, 2001, p. 32) and exclaimed that out of the 1,200 teacher training

programs, no more than 300 would meet a valid set of criteria for such an organization. NCATE was formed in 1952 primarily to oversee the accreditation of programs though a cooperative effort of TEPS, the American Association of Colleges for Teacher Education (AACTE), and the National Association of State Directors of Teacher Education and Certification (NASDTEC) (Angus, 2001).

The way NCATE was governed originally gave more influence to the classroom teacher. This strategy gave representation to classroom teachers through NEA, the state education legal authorities and the schools preparing teachers through the AACTE. In 1954, the 19 members of NCATE consisted of seven AACTE collegiate appointments, six classroom teacher representatives, three college faculty members appointed by the National Board on Accreditation, and the remaining three were represented respectively by the NASDTEC, the Council of Chief State School Officers and the National School Boards Association. NCATE worked to both raise the standard of teaching across the country and to move even further the control of the profession from state level to federal level. In such a scenario, educationalists would have influence nationwide from a more central location. Growth in this direction however was a slow process. From 1954, NCATE's number of accepted institutions grew from about 284 to 342. Additionally, several reconstructions in the makeup of the board reduced again the influence of the classroom teachers. During this time education fell under a great deal of criticism with both the liberal arts departments and the professional schools of education. Included among the criticisms was that entrance and exit requirements for teacher education programs had become low, possibly leading to the increase in state requirements for

initial teaching certificates during the 1950s and 1960s. It was suggested that some classes within education programs were too easy and ironically noted that sufficient scientific evidence was not present to connect several aspects of teacher training to performance in the classroom. The debate was high in the early 1960s as some critics suggested that our entire American school system be replaced by a more European style system of elective secondary schools, while others called for less extreme reform to the current system. One author suggested that the only necessary portion of teacher education was a quality student teaching experience. Many states in the 1950s reviewed thoroughly their teacher certification system. Many changes in the certification systems set the stage for a move away from the centralized system controlled by the educational elite, while teacher voices demanded a say through organizations like the American Federation of Teachers (AFT) (Angus, 2001).

The three major groups, consisting of the professors of education, the professors of liberal arts, and the practicing teachers, continued to struggle for power and influence over how new teachers should be trained even into the late 20th century. This was a significant shift of power as the professional educators held most of the influence in training and certification decisions in the past. Two reports prepared by the Carnegie Task Force on Teaching as a Profession and the Holmes Group attempted to address public concern of the profession (Carnegie Forum on Education, & the Economy, 1986; Lanier, 1986). These reports called for changes such as the elimination of the undergraduate degree in education and insisted that new teachers have a bachelor's degree in their specific field. In the last few decades of the 20th century, mastery

examinations returned as a gateway to the teaching profession in many states. From 1987 to 1994, the percentage of school districts that required a passing score from a state exam grew from about 35% to just fewer than 50% (Angus, 2001).

The history of the teaching profession is largely colored by the struggle between three major groups. These groups include the educator professionals who seek to grow and develop the science behind the profession advocating that teaching is a science to be learned rather than a natural gift, the liberal arts professors who emphasize the need for mastery of content knowledge in the core subjects as a crucial piece of teacher preparation, and the classroom teachers who, through sweat and rigor, experience and share solutions to the most practical problems of the profession. As history has unfolded, temporary victories of one of these groups has shaped seasons of the fields development and highlighted at each juncture an important perspective on what makes a good teacher in American culture today. According to Tobin (2012), the current system of teacher certification reflects a patchwork of interests from multiple stakeholder groups such as governments, educational professionals, the business community and the public. Since the colonial years, the natural progression of teacher certification standards has moved from local control to state and national control. Bales (2006) discussed what she called a tug of war between state and national control of teacher education with organizations such as NCATE, INTASC, and NBPTS putting forth policy agendas to both "teams" to professionalize teacher education.

Among the most recent changes in modern accreditation is the shift from NCATE to the Council for the Accreditation of Educator Preparation (CAEP). In 2012, Ohio

became the first state to sign with CAEP as the new accrediting organization for teacher education and on July 1, 2013, NCATE and TEAC were consolidated into CAEP to take over as the primary organization to fulfill this role. In 2016, CAEP standards will be used exclusively for education preparation accreditation and all legacy standards from NCATE and TEAC will be discontinued ("Council for the Accreditation," 2015). Gateways into the profession and teacher training in the United States have evolved throughout the past century raising the standards of the profession and protecting both qualified educators and students alike. In light of this goal, the following question is merited: do changes in teacher training and certification affect changes in student achievement?

Effects of Training on Teacher Quality

Student achievement is in large part one of the greatest measures of success in our educational system. It can be used as feedback to aid teachers in improving the quality of instructions and in some cases influences teacher retention decisions. When discussing the improvements to any assessment system it is wise to consider the potential benefits that such an endeavor will yield. Because student achievement has come to be perceived as such an influential variable in educational quality the potential effect of teacher training on student achievement will be considered next. Specifically, one common way to address this question in modern research is to explore the effects and correlation of teacher training on student achievement. Using data from 50 states, Darling-Hammond (2000) researched the relationship between teacher quality as well as other inputs from schools related to student achievement, with both qualitative and quantitative methods.

According to her results, measures of teacher preparation and certification correlated with student achievement more than any other variable in the study in the areas of reading and mathematics. This effect holds even after controlling for factors such as poverty and language status. Aaronson, Barrow, and Sander (2007) also found that improved math teacher quality increased student math scores. These results did not appear to be affected much by conditioning variables, and high quality teachers are especially critical for low-ability students.

There are varieties of training and preparation strategies that are recognized in teaching. It has been suggested that the type of teacher certification can explain some of the variation in teacher quality. Two paths of teacher certification are traditional and alternative. In one study by Henry et al. (2014), researchers classified teachers of various entry paths into portals, and explored how teachers classified within these 11 portals (eight traditional, three alternative) impacted student achievement in high school, middle school, and Elementary school across the subject areas of math, reading, science, and high school social studies. Four questions were used to classify teachers into categories.

1. Was the teacher fully qualified, or did they meet all requirements for state licensure?
2. If fully qualified, were the qualifications obtained through classes as part of an undergraduate or graduate degree program or were they part of a program that ended in only a licensure?
3. What was the highest degree held upon first entering the classroom?

4. If the teacher had a degree and was classified as highly qualified, was the degree earned from a public institution within the state, a private university within the state or an out-of-state university?

Based on these questions the 11 mutually exclusive categories created for the study were the following: in-state public undergraduate prepared, in-state public graduate degree prepared, in-state private undergraduate prepared, in-state private graduate degree prepared, out-of-state undergraduate prepared, out-of-state graduate degree prepared, in-state public licensure only, out-of-state licensure only, Teach for America, visiting international faculty, and alternative entry. Additionally, one category was designed to capture all other teachers whom, based on available administrative data, could not be classified into one of the other 10 categories. The total number of teachers in the public schools of North Carolina in 2008-2010 was 100,616 with about 35% attending an in-state public university; of these the sample size for Henry's study was about 30,000. The criteria for a teacher selected for this study was restricted to those with 5 years or less experience, since research suggests that effects of teacher training are generally demonstrated only in the early years of teaching (Goldhaber, Liddle & Theobald, 2013). Next, Henry et al. (2014) measured student achievement by the scores of students on a North Carolina end of grade or end of course exam across the 10 categories of elementary math, science and reading, middle school math, science and reading, and high school math, science, English 1, and social studies. In order to estimate the effectiveness of teachers entering through each of the portals while removing confounding effects, a three level hierarchical linear model was used with students nested in classrooms nested in

schools. The reference group to which all other portals were compared was the in-state public undergraduate group that was compared across the 10 different standardized test measures creating 100 different comparisons. The results found that in-state public graduate prepared teachers were more effective in high school math and the private in-state undergraduate group was less effective in high school and middle school math and elementary science. In addition, in-state private graduate teachers were more effective in high school math than the reference group, while out-of-state undergraduates performed less effectively in half of the 10 categories. Out-of-state graduate degree teachers as well as in-state public licensure only, showed no difference in effectiveness while out-of-state licensure only teachers were less effective in elementary math and reading. Alternative entry teachers were less effective in three comparisons. Teach for America teachers were more effective in seven of 10 comparisons, and teachers entering through the visiting international faculty program were more effective in elementary reading and less effective in high school math (Henry et al., 2014). This study provides evidence that the type of training received by teachers does have an effect on student achievement. This study was a follow up to an earlier study looking at teachers entering the profession during 2007-2008 and with the addition of the portal lateral entry. The previous results found that the out-of-state undergraduate prepared group was generally less effective in elementary schools; Teach for America was generally more effective in middle school math while UNC licensure only teachers were less effective in middle school reading. In the high school arena, the NC private graduate prepared group as well as the Teach for America group were generally more effective than the reference group, while the out-of-

state undergrad prepared group, visiting international faculty, and lateral entry groups were generally less effective (Henry et al., 2010).

These results provide some insight into how student achievement is influence by the path through which a teacher enters the profession. While many states are currently using student achievement data to influence teacher evaluation programs, few use such data to implement or shape policies and procedures related to teacher preparation programs (Bidwell, 2013). If student achievement is influenced by teacher preparation, a next step in improving student achievement is to highlight and continue to develop teacher preparation programs. The instruments used to evaluate potential teachers, whose results are used in license recommendation decisions represent one important factor in the quality of a teacher candidate evaluation programs. High quality, reliable assessments, combined with high standards, continue to produce quality candidates recommended for licensure. Reliability and validity in teacher candidate assessment should be examined using quality methods that aid in justifying the heavy weight placed on these gateways.

The Art of Preparing Teachers

Teaching requires the integration of multiple knowledges and skills applied to unique circumstances involving diverse and unique learners. Teacher education programs are thus faced with the challenge of how to prepare a teacher to face and apply this knowledge base, skill set and character traits to a constantly changing environment, in other words to become "adaptive experts" (Darling-Hammond & Bransford, 2007, p. 391). Naturally, to be successful in such an environment a teacher must enter the field ready to learn and adapt from every encounter, and maintain an up to date knowledge of

both the curriculum and pedagogy, while integrating all this through a character that inspires others to grow and learn. Some elements of curriculum are a direct response to the problems of teaching; others are designed to emphasize and apply a set of standards to act as building blocks by which to construct successful teaching practices. Among the most important is a practicum to practice adapting, decision making, and learning, first under careful supervision and in an environment build to provide feedback (Darling-Hammond & Bransford, 2007). Historically, large emphasize was placed on content knowledge as being one of the most important aspects of teaching (Shulman, 1986). Up until about the mid-1990s, research on teachers was mostly limited to observing correlations between practices and student achievement. It was around this time that a shift occurred in perception of what makes a good teacher. The emphasis begin to shift to characteristics of the teacher such as thought process in decision making, lesson planning, and beliefs rather than simply observable actions. The idea of one-directional causality and assumed linearity of relationships between student success and teacher effect begin to be questioned (Fang, 1996). Dispositions, often referring to as the inherent characteristics of a teacher, stem from beliefs, attitudes, and experience. Interest in these dispositions eventually led to NCATE requiring their evaluation (Almerico et al., 2011). Another struggle preparation programs face is to establish a clear definition to accurately evaluate them. As one author mentions, "any evaluation process must rely on clearly defined constructs that cannot be interpreted in open-ended ways to suit the subjective biases of the evaluator" (Damon, 2007, p.1). Ingersoll, Merrill and May (2014) found that a teacher candidate's training in teaching methods and pedagogy along

with observations, feedback, and experience in the classroom or lack thereof correlated with whether or not the teacher would leave the profession after a year. Those with more of this type of training were far less likely to leave after a year than those without. Considerations of this type made their way into both educational training programs and teacher assessments with designs to measure changes in teacher growth and dispositions over time.

Once instruments were created to assess the skill needed, they were put into practice and examined. A teacher's pedagogy/psychological knowledge (PPK) was measured by an instrument using 39 multiple choice items, short answer items, and video items and was intended to demonstrate knowledge across the spectrum of teaching including sub-dimensions such as classroom management, teaching methods, classroom assessments, learning processes, and individual characteristics. Evidence was presented for the validity of this assessment through both statistical procedures and expert opinions (Voss et al., 2011). The Performance Assessment for California Teachers (PACT), which scored teacher categories such as planning, instruction, assessment and academic language was assessed and found to be a valid instrument for assessing teacher competence as it affects licensure decisions (Pecheone & Chung, 2006). Another instrument based on the four dimensions presented by Danielson (2011) including planning and preparation, the classroom environment, instruction and professional responsibilities, was examined for reliability and validity, comparing scores of three types of raters including the supervising teacher, onsite evaluator and the student teacher (Benjamin, 2002). EdTPA is an assessment system currently used throughout the nation,

which attempts to emphasize, measure, and support subject specific and general knowledge of teacher planning, instruction, and assessment (Pearson Education, 2016). While there is much overlap in knowledge required to succeed in teaching, the specifics are still evolving to train and test teachers in ways that make the most impact possible on the education profession and the development of today's students.

Validity and Reliability

For any instrument designed to evaluate at a high-stakes level, reliability and validity are important factors in the discussion. If you consider a bullseye, reliability can be thought of as precision and validity as accuracy. They are not dependent on one another and an instrument can have neither, both or reliability without validity. Specifically, when considering the reliability and validity of an assessment instrument, this idea is expanded. Reliability of an assessment can be thought of as whether an assessment would produce the same or similar results when administered to the same or a similar population. This would look like hitting the bullseye in the same place whether or not it is in the center. Where validity considers whether the instrument we use actually measures what is intended. This would look like hitting the center of the bullseye even if the shots are not necessarily grouped together.

Wolming and Wikström (2010) explored the changes in definitions given to validity over recent years, and argued that it has become more broad, and that there lacks evidence for a unified validity argument. Furthermore, that while practice has often kept up with theory in how research is designed it often deviates when these designs are

carried out in practice. Validity put into practice and methods of producing specific validity evidence is still in need of guidance.

There are of course several pieces to both reliability and validity and when they are applied to research it is important to consider which aspects apply most to the specific topic at hand and with appropriate emphasis. According to Moskal and Leydens (2000), there are three primary categories of validity and two major categories of reliability that are relevant when dealing with the assessment of student performance. The three aspects of validity include content validity, construct validity, and criterion validity while reliability is made up of inter-rater and intra-rater reliability.

Validity is a word we often use to describe how well something measures what it intends. Content validity can be thought of as how well the content within an assessment reflects the concept being measured and furthermore, avoids measuring accidental or unintentional variables whose interpretation could interfere with the true purpose of the instrument. For example, if we are trying to measure knowledge of history and yet use questions on the assessment that require a high level of English proficiency to understand, unless all the students are highly proficient in English, the assessment can end up measuring two constructs instead of only the one intended. In such a case, examinees would only achieve a correct response if they have sufficient proficiency in both history and English. In this example, a lack of knowledge in either history or English would likely cause an incorrect response. An evaluator may incorrectly assume that history proficiency is low when in fact the instrument has lost its ability to measure independently this construct. Content validity is also concerned with whether an

assessment samples fully the domain of what is being measured. For example, results of a math test comprised mostly of addition problems cannot be used to generalize about math ability as a whole. For results of an assessment to be able to adequately inform regarding all aspects of basic mathematics, sufficient examples of each important part must be included on the assessment (Moskal & Leydens, 2000).

A second aspect of validity, called construct validity, differs from content validity in that it speaks to how well the assessment actually measures its intention. An example where construct validity is not present would be if an assessment claimed to measure understanding of a concept or thought process as only implied through a correct answer, where no opportunity is given for the examinee to demonstrate how they reached the conclusions they recorded. In several types of assessments, it is possible to score a correct answer on a question while failing to conceptually understand the process. This type of problem might be addressed through including items that address the process. For example, requesting a student justify their response and reflect this in the scoring rubric through partial credit. At this point, claims that both knowledge and reasoning are being measured become valid.

Criterion validity relates to how well the skills tested on the assessment correlate to success in the profession or task. Specifically, for an assessment required for licensure, this aspect of validity could be measured by observing the relationship of success on the instrument with success in the teaching profession (Moskal & Leydens, 2000). Criterion validity would be possible to measure by correlating scores on exit

assessments with performance scores on professional evaluations or student achievements.

Newer developments in the definition of validity have become popular and relate more closely to the organization of evidence to support validity. Kane (2013) argued that the claims we make about an examinee based on a test score extend far beyond performance even to proficiency in a specific area, and evaluating validity can be thought of as evaluating the plausibility of these claims. Cook, Brydges, Ginsburg, and Hatala, (2015) summarized Kane's framework in considering four aspects of validity including scoring, generalization, extrapolation, and implications. Scoring has to do with factors such as item performance and characteristics such as difficulty and discrimination, rubrics, and reliability. The second aspect of Kane's framework referred to what extent the results can be generalized. Measures of this include sample size, selection criteria, or types and amount of missing data. In this research, the complete sample of 3 years of student performance data was used; however, missing data especially in small departments as well as limited availability of data for candidates who did not pass the assessment limit generalizability to the larger programs and more specifically to completers of those programs. Extrapolation expands this generalization to how strongly these results apply to not only the testing world but the real world. This type of validity would be supported by evidence that test performance correlates to success in the field. While information on teacher candidates who continue their profession in NC are currently collected, the collection of data from teacher candidates who continue their careers outside of NC would provide the majority of the data needed to evaluate such a

claim. Extrapolation also speaks generally to the potential factor structure of an instrument or assessment system which as applied to the assessment system at UNCG is developed further in a later section. The implications lead researchers to ask the following questions: what does a score on a particular instrument mean? What happens to those who pass? What happens to those who don't? What are the consequences of each result?

The North Carolina Board of Education charged the NCPTS Commission to create new teaching standards, combining the older standards with the new mission of the following: "Every public school student will graduate from high school, globally competitive for work and postsecondary education and prepared for life in the 21st century" (North Carolina Professional Teaching Standards, 2011, p. 1). Several of the factors used in the UNCG student assessment system were retained in the final structural equation model can be mapped onto these standards as well as the InTASC standards (Assessment & Support Consortium, 2011). For example, Standard Three of the NCPTS requires that teachers know the content they teach which is aligned with and demonstrated by Evidence Two in the UNCG assessment system, which requires the student to produce and in-depth inquiry project about a topic in their academic field. Also, NCPTS standard one requires a teacher to demonstrate leadership, which is demonstrated by evidence six in the UNCG assessment system. A more complete mapping of the factors used in the final structural equation model onto state and federal teaching standards can be found in the appendix.

Reliability can be broken down into several parts. An exam is considered to have high reliability if several administrations under the same circumstances, would yield similar results. Moskal and Leydens (2000) break down reliability in regard to student assessment into inter-rater and intra-rater reliability. Inter-rater reliability is high when there is a higher degree of agreement of scores where the same student is evaluated by multiple evaluators. One way to reduce variation is by presenting all evaluators with a clearly defined and well-constructed rubric by which to grade each exam. If a score of "3," for example, is clearly defined and thus is interpreted similarly in the minds of multiple evaluators, consistent results are more likely to be produced. This may not solve the problem completely but evaluator training can take steps to increase confidence that each score represents the same level of competency across evaluators. Another method that can be used to increase inter-rater reliability is to use anchor papers which might represent a model paper of a specific level of each criterion on a shared rubric. Depending on the complexity of the assessment inter-rater reliability can be addressed by reporting a percentage of agreement, a *t*-test, or ANOVA for simple data, or methods involving structural equation modeling for more complex data sets where levels of invariance between evaluator groups can be tested across the model (Moskal & Leydens, 2000).

The second category of reliability is intra-rater reliability which can be thought of as how robust a raters scoring would be to external factors such as fatigue, mood, pressure, or student bias. A rater for example might score students differently depending on their mood, for example, they are under pressure to meet a deadline. Also, raters

could be more prone to giving a successful student a high score and/or a struggling student might yield a more critical eye for grading, a phenomenon known as the "halo effect" (Thorndike, 1920). This problem can be addressed by having evaluators grade de-identified papers, or by using clearly defined rubrics where raters can revisit exactly what each score category should look like by description or example (Moskal & Leydens, 2000). Open discussion during evaluator training sessions about these issues also has great potential for minimizing their negative effects. Although some of these factors are internal to the rater themselves, they can often be addressed through training and increasing awareness of such risks to reliability. Potential for such risks might be explored through carefully worded surveys to look for hints of changes in grading under different circumstances. Depending on each rater's self-awareness, simply being made aware of the risk to reliability may encourage them to alter factors under their control to reduce these risks.

Consistency between forms of an assessment is another aspect of reliability. If no examinees receive the same instrument it is important to provide evidence that these forms are equally capable of measuring the desired construct. Some statistical methods exist for dealing with this kind of reliability in determining levels of re-test reliability, alternate forms, or split-halves reliability. All of these methods are used to provide evidence that different forms consistently measure the same construct (Carmines & Zeller, 1979).

All relevant aspects of both reliability and validity merit discussion for any assessment, and limitations should be openly acknowledged. On a reliable assessment

student scores should be consistent despite factors such as when the student took the exam, when the scores were graded and recorded and the identity of the evaluator. If these criteria do not hold, student scores are at least in part dependent on factors unrelated to the purpose of the instrument (Moskal & Leydens, 2000).

Research in Education

Farkas et al. (1997) surveyed 900 teacher educators through phone surveys. The sample represented successful responses from 5,324 teacher educators who were randomly sampled from a pool of 34,000 that included deans, chairpersons, and faculty members. Within the summary of findings, it was concluded that many felt detached from today's school and had doubts that they had adequately prepared students for success. It is further stated by Raths and Lyman (2003) that there exists a lack of respect for the work put forth by faculty of teacher education to make summative evaluations of new teachers. While there are several reasons associated with this lack of confidence, part at least is attributed to lack of evaluator training and questionable measurement instruments (Barrett, 1986). If instruments used as gateways for licensure recommendations are not reliable, the results are not necessarily due to the factors we anticipate. This can cause interpretation of results to be misleading, incomplete, and ultimately untrustworthy. For example, if inconsistency exists in rater scores whether between raters or within raters, then final scores are attributed at least in part to the evaluator assigned to the student or the circumstances by which they are evaluated rather than the candidate's competence and ability. The ramifications of untrustworthy results can affect both public and professional views of the entire evaluation system.

Alternatively, providing evidence for the reliability and validity of student teacher evaluation instruments can be used to reinforce confidence in both the license and the candidates who receive them.

InTASC, created in 1987, represents both state education agencies as well as national educational organizations. InTASC has created a set of standards intended to reflect attributes of quality teachers across subject areas with the belief that "an effective teacher must be able to integrate content knowledge with the specific strengths and needs of students to assure that all students learn and perform at high levels" (Interstate Teacher Assessment, 2015, p. 1). These standards are commonly used as a guide when constructing and implementing student teacher assessment instruments. While the InTASC standards are useful for demonstrating types of problems such as planning, instruction, classroom management, and assessment, which are associated with a group of student teachers, they may not capture problems associated with the personality of the student teachers such as time management and personal organization (Jaus, 1999). Simpson (2004) researched the contribution of exit portfolios in assessing student teachers. She found that neither demographics nor other assessments such as student teacher grades, or Praxis results seemed to influence the results of exit portfolios and thus concluded that these assessments provided a unique contribution to the assessment of student teacher candidates. Bates and Burbank (2008) discovered that under the accountability of No Child Left Behind, university supervisors struggled with a number of factors in student evaluation ranging from inconsistencies between formal and informal feedback given to students, to perceptions of candidate success influencing

whether feedback was tailored toward individual needs or strictly based on standards. In cases where this was observed, candidates not perceived as successful were often provided with feedback based on standards, where successful students often caught the eye of evaluators and were given more personalized feedback.

One recent study by Miles and House (2015) called into question the reliability and validity of student teacher evaluations based on findings that results on such instruments can be influenced by factors outside of the students control. Several of these factors include class size, course type (elective or required), professor gender, and course grades. These variables seem to influence results on student teacher evaluations however have little value in assessing teacher effectiveness. Based on over 30,000 student evaluations of 255 professors it was found that evaluations are most likely to be scored highly in small elective classes and least likely to be scored highly in large required classes with female instructors. In addition, evidence was found that the expectation of high course grades may influence higher scores on evaluations.

With confidence in educational preparation programs becoming a concern, practical and specific insight into the details of teacher preparation assessment instruments is both timely and relevant. Through statistical research methods, we can confirm strengths and highlight specific potential problems in these instruments, giving education departments new tools and confidence to continue building teacher preparation programs and serving teacher candidates.

Other Research

While structural equation modeling applied specifically to address the issues involved in a teacher candidate assessment system is new, the application of these analyses to address other similar types of assessments in both education and other fields are common. Konkolÿ Thege, Kovács, and Balog (2014) used a bi-factor model to describe the Posttraumatic Growth Inventory assessment, which models how 21 items contributed both to a general factor (labeled posttraumatic growth) and a specific factor each item was believed to represent (including relating to others, new possibilities, personal strength, spiritual change and appreciation of life). Another example of a bi-factor application is Xie et al. (2012). They described a patient's level of distress or pain as a general factor and the specific factors as depression and anxiety when modeling the Hospital anxiety and depression scale. This research also used results from the bi-factor analysis to calculate the test information function that demonstrates at what level of distress of patients in general is the assessment most informative to discern accurate measures of distress. Yang et al. (2013) compared the fit of a bi-factor model with a single factor and a five-factor model showing the bi-factor model gave the best model fit to describe the relationship of the factors in the Acute Stress Response Scale. Using the results, a test information curve was produced describing how the test preformed across various levels severity of stress response. Finch (2005) used a MIMIC CFA to detect differential item functioning, and compared his method with other DIF detection methods such as Mantel-Haenszel, SIBTEST, and the IRT Likelihood ratio. Results showed that his method was effective at detecting DIF with a large number of items and with data best

described by the two parameter logistic model; however, it was less effective with fewer items dealing with three parameter data. To explore the major differences in the methodologies for detecting DIF the MIMIC model estimates direct and indirect effects on a grouping variable. After controlling for group differences by regressing the latent trait onto the grouping variable, any remaining direct effects on the item in question from the grouping variable demonstrates item DIF. This is measured in this case by observing the modification indices to see if after controlling for the effect of the latent trait on the grouping variable, there is any significant loading onto the item itself. If DIF exists, this item is flagged allowing the loading from the item to the grouping variable to be estimated and the model is re-run (Finch, 2005). SIBTEST on the other hand is a nonparametric method that estimates both the presence and amount of DIF in targeted items between a reference and a focal group. This procedure compares the performance of examinees of similar ability between the two groups, accounting for possible differences in the ability distributions between the two groups. It also compares a valid subtest assumed to contain items without DIF to a questionable subtest that contains the items suspected for DIF and compares performances on the two sub-tests. This difference is used to estimate the presence and amount of DIF between items (Bolt, 2000). Woods (2009) later extended this research, exploring further the sample size requirements and accuracy of the MIMIC model for detecting DIF and added support for the approach. Oliden (2011) evaluated the type 1 error and power of using a multigroup CFA invariance test to detect both uniform and non-uniform DIF. Results were encouraging in both cases when DIF detection was determined by comparing a chi-

squared difference test (both Bonferroni corrected and uncorrected) and a difference in Comparative Fit Index (CFI) values to flag items with potential DIF. Chang, Huang, and Tsai (2015) found that using a multiple-group categorical confirmatory factor analysis (MCCFA) with minimum free baseline approach was also an effective strategy for detecting DIF in polytomous items. Fukuhara and Kamata (2011) demonstrated a method for DIF detection on testlet based data using a bi-factor multidimensional IRT model and as this model took testlet effects into account it was found to have better estimates of DIF, as well as higher detection rates, than more traditional IRT DIF models.

Quantitative Research of Teacher Assessment

This research includes a fit assessment of the teacher education data available using multiple a confirmatory factor analyses. The purpose of a factor analysis is to describe the relationships between observable variables and latent (unobservable) traits. This assumes that ability in a specific area like content knowledge, which we can't observe, will predict the same person's ability to score well on an item related to content knowledge, which we can observe. This relationship will not be perfect but one of the advantages of structural equation modeling is its ability to account for this imperfection through the use of error terms, that would average to zero. In this case we take what we can observe, which includes evaluator responses to several items, and use this information to make inferences about the latent ability or abilities of the teacher candidates. As the assessment system is broken up into various sections that as a whole are used to make a summative judgement, it follows that the instrument attempts to measure several sub-factors that have some relationship to each other. This relationship,

depending on the model, could be a series of unexplained correlations, factors that are present and independent once the commonality shared by the general factor is accounted for or the correlations between the factors may be fully explained by the general factor. In other words, we might say that the ability to plan a lesson and the ability to successfully evaluate students might be independent traits if one could theoretically remove the commonality that is shared between the two because someone is a good teacher or we might say that being a good teacher accounts for all that is shared between these two skills. A bi-factor structural equation model attempts to explain the covariance of the data using the general factor and a specific factor that each item is associated with representing the remaining variance after the general factor is accounted for. A higher order model can be viewed as all the sub-factors such as teacher dispositions, or growth or the ability to plan a lesson are simply pieces that ultimately make up being a good teacher but do not make sense to consider outside of this framework. Reasonable models to be considered for describing such data might include several possibilities including the following:

1. a single factor model if you assumed that every item was best described only as a piece of the larger test,
2. a correlated factors model if you wanted to display the data as representing multiple factors but wanted to allow them to correlate while giving no attempt to model the structure of this correlation or to measure a single common construct,

3. a bi-factor model if you wanted each item to be able to load on both a general factor representing the whole assessment and a specific factor representing a smaller sub-factor,
4. a multiple factor orthogonal model with multiple latent factors where latent traits are not allowed to correlate and are assumed to independently measure different aspect of the larger exam, or
5. a higher order model if you believed that sub factors were nested within the larger factor and observable data was a result of ability on a latent trait which was also influenced by a more general latent trait.

Which one is most appropriate depends largely on how the assessment is conceptualized. Structural equation modeling can help us here by allowing us to see which of these stories, the data support as measure by multiple fit indices. The fit does not guarantee the model is correct, but only lends support to a model that is already considered to make sense. Although bi-factor models have not always been a common method used to describe this type of data, several instruments used today in both psychology and related fields are implemented and interpreted based on the assumption that a general construct is measured by several closely related domains (Wiesner & Schanding, 2013). Structural equation modeling is a technique used for specifying models of linear relationships between observed variables and latent (unobserved variables) and the model itself is a hypothesized pattern of relationships between these variables (MacCallum & Austin, 2000). Since the 1970s the use of structural equation modeling has grown rapidly. As advances in technology and software have made the

practical application of SEM more feasible, the literature has filled with examples of its use in psychology and related fields (Bentler, 1986; Tremblay & Gardner, 1996). Structural equation modeling uses a variety of techniques depending on the data to fit a model. One common technique for continuous data is to compare the covariances of a group of variables from a data set and then attempt to minimize the difference between the model covariance matrix and the data covariance matrix assuming conditions specified by the model. If the difference is small, the model is considered to have good fit or in other words may accurately represent the relationships of the variables. If the model does not fit, it is probably not a good description of these relationships. Good fit does not ensure true representation. This is why it is important in SEM to have some justification for the specification of your model before fitting it to a CFA.

Following the model fit alone can lead in a direction that is neither practical nor useful. *Parsimony* refers to the idea that other things being equal preference should be given to the simpler model. Several fit indices used in CFA account for model complexity and add a penalty to the result for overly complex models to varying degrees. This allows these indices to intentionally favor the simpler solution if other factors are similar, which can be very useful to researchers. Some examples on indices that are adjusted for model complexity include the Akaike Information Criterion (AIC; Akaike, 1974) and the Bayseian Information Criterion (BIC; Schwarz, 1978) which are fit using maximum likelihood, and more appropriate for non-normal data the Tuker-Lewis Index (TLC; Tucker & Lewis, 1973). Unlike many common fit indices, the TLC was also found to be independent of sample size in a study that used both simulation and real data

(Marsh, Balla, & McDonald, 1988). Other techniques in SEM are designed for other types of data; for example, ordinal data can be modeled using CFA but not in the same way. Methods such as weighted least squares (WLS) and diagonal weighted least squares (DWLS) differ from maximum likelihood estimation techniques in that it does not rely on normality within the data. Ordinal or categorical estimation techniques assume that a continuous latent variable is influenced by the observed ordinal variables and a polychoric correlation matrix represents the estimates of the correlations between these latent continuous variable. Flora and Curran (2004) demonstrated that estimation of polychoric correlations were robust to moderate violation of normality within this underlying continuous latent variable, and that robust WLS performed well in estimating CFA models for ordinal and categorical data and WLS performed adequately only when sample sizes were large.

SEM can measure reliability, test and item quality, and the fit of the assumed structure of the data. The data fit to the SEM models are student achievement scores from about 27 programs, across three primary assessment instruments made up of about 12 constructs comprised of 78 assessment criteria items, across 3 academic years. In some cases, these scores were averaged across multiple evaluator scores.

In CFA, a model is not created based on data fit and modification indices. A model developed in this way would not necessarily reflect reality or make practical sense. Instead, models that are built *a priori* from theory or purpose are tested, which are believed to fit reality, for statistical fit. Whereas exploratory factor analysis can be useful to give you some idea of a model when you don't know what may or may not fit reality, a

model that fits the data better than any other does not necessarily reflect reality best, but can be used to raise questions about the patterns of covariance within the items.

Once a model is developed based on theory and confirmed with goodness-of-fit tests, inter-rater reliability can be tested. Data represents an average of the scores of multiple raters in some programs, and therefore it is important to consider whether multiple raters score students consistently. Because the data is composed of items that are assumed to make up several constructs, it is important to take into account this mapping when asking questions about differences in groups. A multi-group CFA model can compare the scores from the different types of evaluators across these latent factors. For the purposes of this study, evaluators have been classified as either an onsite teacher evaluator (OSTE) who is employed by the school at which the candidate fulfills their student teaching experience or a school supervisor employed by UNCG. The teacher candidates who were evaluated are the same in each group, so it is important to discuss the issue of independence. Since the scores represent the same people in each group the analysis can be thought of as similar to a repeated measures analysis over multiple factors where instead of dealing with time points measuring how the same groups changed over time, instead we measure how the same people were evaluated near the same time point by different raters. A change over time would show a difference in the raters as opposed to a change as a result of time or training. One example of this type of analysis was performed by Schaie, Maitland, Willis and Intrieri (1998), who examined the equivalence of the factor structure of the psychometric ability tests over the course of 7 years. For this study, 984 persons were tested twice, approximately 7 years apart, on 20 tests

making up six factors. To account for the dependence of groups the factor structure was extended to a repeated measures multi group factor model for panel data. Ployhart and Oswald (2004) also discussed repeated measures type data and mean and covariance structure analysis (MACS) is appropriate. MACS tests simultaneously whether the measure is equivalent between groups, if the latent construct shows differences between groups with regard to the variance and covariance, and also whether differences exist between latent means. Deng and Yuan (2015) also develop a method for multi-group structural equation modeling that does not require specification of between group relationships. When correlations between groups exist and those are ignored, the model is mis-specified and this can run the risk of invalidating results. Pentz and Chou (1994) used a longitudinal SEM model to address the issue of invariance where samples contained the same subjects across groups with time as the variable.

Factorial invariance is critical when discussing the comparisons of latent means across groups. This refers to the extent to which the CFA model holds for both groups in the same way. If invariance holds, differences in latent means can be interpreted as true differences and not due to differences in the strength of loading or intercepts between the groups. Invariance was tested in multiple stages beginning with the basic model structure for each group and then testing more and more strict levels of invariance to see if the models that represent each group are the same. A method for testing invariance includes up to four levels of invariance testing to establish invariance across groups then two additional levels of invariance specifically related to the latent variable, which are often the tests of interest to the researcher. The four common tests for invariance are in order;

configural invariance, metric invariance, scalar invariance, and strict invariance. Configural invariance considers whether the same basic pattern of loading and latent variable holds for all groups. If configural invariance holds, then metric invariance tests whether both groups have the same loadings of the observed variables on the latent factor. If metric invariance holds the model is considered to have established weak invariance, and causes for differences in the latent variable are then isolated to true differences or differences in the intercepts.

To establish strong or scalar invariance, the next step is to test for equality across groups of both factor loadings and intercepts, or thresholds. If strong invariance is established this is considered sufficient evidence to interpret differences in the mean structure of the latent variable; however, one more test could be done which established strict invariance. Strict invariance holds if the loadings, the intercepts (or thresholds) and also the residual variances are the same across groups. Each of these tests is performed by comparing the chi-squared difference, or some other established index or combination thereof, between the models. If the stricter model does not statistically have a worse fit than a less strict model, then invariance holds for that level of strictness (Oliden, 2011). Van de Schoot, Lugtig and Hox (2012) suggested that factor loadings and intercepts be tested separately with the other free to vary before testing for strong invariance. Millsap and Yun-Tein (2004) compared how factor invariance was tested in the two popular software packages Lisrel 8.52 (Jöreskog & Sörbom, 1996) and Mplus 2.12 (Muthén & Muthén, 2010). The results showed that Lisrel estimates thresholds for the combined group and then uses the same thresholds to estimate other variables for each group

separately, treating them as fixed parameters. Since model specification is different in each software, fit results were also different in some cases.

To evaluate the performance of the items and the test across the ability scale, item loadings on the general factor, or alternatively specific factors if no general factors is retained, and thresholds will be examined. In several cases depending on the model retained an IRT model corresponds to the SEM model where item discriminations are related to the loadings and item difficulties are related to the thresholds by a specific conversion formula. Assuming the single factor model was retained, the corresponding IRT model would be Samejima's (1969) graded response model. If the uncorrelated six-factor model were retained the corresponding IRT model would be a multi-dimensional polytomous, IRT model for ordinal data such as proposed by Bacci, Bartolucci, and Gnaldi (2014).

While the loadings of the CFA model determine how the items perform in measuring their intended factor(s), the thresholds illustrated the difficulty of each item or more specifically the amount of the latent ability required to have the best chance to move from one specific score category to the next. In a simple CFA each observed variable "y" can be written as a linear relationship to the latent (unobservable) variable "η" as $y_{j=}\lambda_j\eta + e_j$ where λ is a calculated loading (also called the slope parameter) and e is and error term (or residual) accounting for the inability of the latent variable to completely predict the observable variable. The error terms will have a mean of zero and are assumed to be independent of the latent variable. For the parameter logistic model,

the formulas to convert CFA model loading parameters to IRT discrimination parameters is the following:

$$a_j = \lambda_j(1 - \lambda_j^2\psi)^{-1/2}\sigma_{nn}^{1/2} \qquad (2)$$

where a is the discrimination parameter, λ is the loading onto the general factor, ψ is the error variance for the latent variable and σ_{nn} is the variance of the latent trait. In a DIF analysis it is assumed an outside trait Z that influences the latent trait in a specific way and if DIF exists will model the difficulty parameter(s) differently depending on group membership. This alters slightly the CFA model to

$$y_{j=\lambda_j}n + B_jZ_k + e_j \qquad (3)$$

if B_jZ_k is non-zero this is evidence for the existence of DIF on that item. This also adds to the model a formula to convert the item threshold τ_j to an IRT difficulty parameter using the following formula

$$b_{jk} = [(\tau_j - B_jZ_k)\lambda_j^{-1} - \mu_n]\sigma_{nn}^{-1/2} \qquad (4)$$

where μ_n is the mean of the latent variable. These formulas have produced IRT item parameters that have shown to be very close to IRT parameters estimates derived through IRT software such as Multilog (MacIntosh & Hashim, 2003; Muthén, Kao, & Burstein, 1991). These formulas simplify when the solution is standardized, making the latent mean zero and the variance equal to one.

Using the item loadings and thresholds, Mplus can calculate directly information and characteristic curves on both the item and test level with regard to any factor. The information curves demonstrate the capacity of the items or instrument to accurately discern the ability of teacher candidates in the factor of interest measured across different levels of that ability. Depending on the model, information could be calculated for the whole assessment representing the ability "teaching capacity," or it might make more sense to have an information curve for each factor separately. The shape of the item information curve will help determine what ability level(s) of students each item can discern across the ability spectrum. The item and test information curves are one of the major advantages of using an IRT model as opposed to simply doing item analysis using only classical test theory. Whereas the latter can give an idea of the instruments reliability at the mean level of ability, IRT can show how that reliability changes for candidates of varying the ability levels. The test information curve is the sum of all item information curves (IIC) across the ability scale and the test information curve is considered in terms of the goals of the assessment.

For example, if there is little interest in distinguishing between an excellent student and a good student as they both pass, but instead a lower cut score determines whether a student passes or fails, then a test where the information curve is highest at the ability level near that cut score would be desired. Information at other levels of ability may or may not be a concern. If the highest point on the information curve is both large in magnitude and near the cut score, a student whose true ability would earn them a passing score if the assessment were perfect, is less likely to be misclassified even if their

ability is very close to the cut score because the precision of the instrument at this ability level is very high. Because information is inversely proportional to the standard error, this can be thought of as minimizing the number of people whose range of standard errors fall on both sides of the cut score. High information near the cut score means the standard error is lower for people near the cut score so the distance from the cut score that includes people with a standard error on both sides is small. In some assessments there may be benefit to having high information in multiple points or regions across the ability spectrum as well. An information curve can be calculated for each item, so it is possible to draw conclusions about how well individual items are aligned with the goals of the larger assessment. If an item information is high near a test level cut score of interest this item can be said to have high precision at this ability level and high contribution to the purpose of the instrument. Alternatively, if there is nowhere on the ability scale that an item has high information, or if the only places where high information is present is far from any meaningful cut score, it could be concluded that this item may not add value to the overall instrument and the items removal should be considered.

The results of these analyses should provide several insights as to how the instruments are performing, and how differences in groups may influence evaluation scores. Some aspects of reliability and test quality are not as straightforward to measure, because the areas of concern are not easily quantified. In such cases, it is possible to expand the methodologies used to include qualitative techniques as well as quantitative in

order to add a new layer of richness to the data, which can enhance the quality and type of conclusions that can be drawn.

Qualitative Research on Teacher Education Assessment

While quantitative measures can provide useful evidence that can be used to make summary statements about instruments and items, often the question of why and how are better answered through a different approach. Internal consistency within evaluators is another type of reliability that addresses whether evaluators tend to be consistent in grading teacher candidates despite their own circumstances. For example, could the score a student receives from some evaluators be dependent on factors like mood, time of day, or circumstances related to the evaluator? One way to approach answering this question might be to compare the scores of several evaluators that scored the same student and simply look for consistency in trends. Does one evaluator stay lower than the others for a while and then in some places peak above the rest? Unfortunately, with this data there is little consistency in who grades multiple students as the same OSTE and supervising teacher do not necessarily grade all students within a group, especially the OSTE. Furthermore, the OSTE and supervising teacher are usually the only ones that grade an individual student. However, this could be explored through carefully worded surveys, and interviews recording grading habits of evaluators. This research is primary quantitative in nature; however, some aspects of qualitative research and its potential benefits to the evaluation of teacher candidate assessment will be addressed briefly for future researchers to consider.

Summary

The culture of education is shaped by a long history of reform and power struggles, as multiple groups with various agendas seek to serve students through education. In the context of modern research there exists evidence that the way we prepare student teachers does influence student learning, which leaves for education faculty and policy makers alike the task of discovering and putting into practice the best methods of preparing educators. One contribution this research provides to the continuous improvement of teacher preparation is to examine psychometrically the student teacher training process, specifically the exiting assessments. Providing evidence for the reliability and validity of assessments used in determining licensure recommendation decisions, as well as for the quality of both the items and the assessment adds value to teaching licenses in the eyes of not only educators but employers and the public. Whereas research has applied advanced analyses to current assessment data, this research explores further specific aspects of the assessment system of teacher candidates at the UNCG. The purpose of this research is to use a set of statistical methods within SEM, to provide evidence for assessment quality as well as highlight potential problems in the spirit of continuous improvement. These findings will serve the specific needs of UNCG and further discussions at other universities seeking to improve their own system of preparing student teachers to a higher level of mastery.

CHAPTER III

METHODOLOGY

Introduction

Student teachers are evaluated based on a number of criteria at UNCG. Among these criteria are the evidences, dispositions, and TGAP instruments. These instruments attempt to measure teacher skills and attitudes at various points throughout a student teacher's progress toward graduation. As a part of the continuous improvement to the assessment system at UNCG, this research will include a psychometric evaluation of the evidences, TGAP, and disposition instruments using structural equation modeling. This research will add to the literature a fresh application of SEM to teacher candidate evaluation and educational research.

Conversations with university employees, stakeholders, and representatives of the UNCG School of Education assessment process as well as experience working with this data have all influenced the outline and goals of this research. Following are the research questions that will help provide focus to this study along with the specific methodologies that will be implemented to address and answer each question.

Research Questions

1. What CFA model best represents the structure of this assessment system given the data?

2. What can we tell about the items and the assessment when a model structure is imposed on the data?
3. Do differences exist in the scores given by the supervising teacher and the cooperating teacher? If so, what is the nature of this difference in terms of the factor structure?
4. Do differences exist in the way items are interpreted and scored depending on whether the teacher candidate is in an elementary education program or in a middle grade/secondary program?

The Data

The data available for the three assessment instruments include scores for all undergraduate and graduate initial licensure teaching completers from the academic years 2011-2012, 2012-2013, and 2013-2014. Although data was available for advanced teaching completers and completers from non-teaching programs in each of these academic years, only the initial licensure assessment system was standardized to include an evaluation for six evidences (comprised of many criteria), nine CDAP criteria, and 18 TGAP criteria. While items exist which are common to most advanced programs, there is a large amount of variability in items between advanced programs, and often a very small numbers of completers in some of these programs. This evaluation will focus on analyzing data from initial teaching programs. It is important to note that the scale is different for the evidences and the CDAP and TGAP instruments. Also the CDAP and TGAP assessments are scored over three different time points throughout a student's progression and evidences are scored only once at the end. For each of the evidence

criteria, students can be scored from one to three with averages of multiple evaluators allowed, and for each CDAP or TGAP criterion students can be scored from one to six with averages of multiple evaluators allowed. It is important to note two things. First, that non-integer scores are present in the data, and for the TGAP and CDAP make up about 17% of the data, where in the evidences it is about 2.1%. If half intervals including 1.5, 2.5, 3.5, 4.5, and 5.5 are included, the items that are decimals not included in this list drops to 3.3% for the CDAP and TGAP and 0.01% for the evidences. Second, a score of one on any evidence criteria is grounds for failure to complete the program, as is a score of one or two on any CDAP or TGAP criterion. Because the data is comprised of completers, scores of one and two do not belong in the final assessment data, but they are present (scores of 1's and 2's act as a flag for earlier time points and can be given without failing a student). Table 2 summarizes the number of students who completed in each program in each academic year.

Table 2

Number of Completers by Program

	Academic Year			
	2011-12	2012-13	2013-14	Total
Art	8	17	13	38
ASL	0	2	0	2
BK	3	6	31	40
Dance	5	1	1	7
DHH	3	2	1	6
Elem / GC	23	24	15	62
Elem	160	116	93	369
English	20	14	11	45
ENRICH	18	0	0	18
General Curriculum	9	21	19	49
Health and PE	11	18	8	37
Latin	1	2	1	4
Math	3	2	3	8
Middle Grades	18	19	16	53
Music	27	22	25	74
NCTEACH	11	6	6	23
Science	2	3	3	8
Spanish	1	2	3	6
Social Studies	22	12	19	53
Theater	6	3	8	17
Undergrad Total	351	292	276	919
Elem GRAD	20	10	16	46
ESL GRAD	12	9	3	24
GC PAIL GRAD	0	15	3	18
Latin GRAD	0	1	1	2
Middle Grades GRAD	0	2	4	6
Science GRAD	0	0	2	2
Social Studies GRAD	2	1	3	6
Graduate Total	34	38	32	104
Grand Total	385	330	308	1023

**Note*. About 844 candidates contain no missing data.

For each completer score, ratings were given by at least one evaluator. When multiple evaluators are present, the average was recorded as the final score for that criterion. In several cases of multiple evaluators, a non-integer score results. However, when half scores are considered, the remaining non-integer scores make up less than 5% of the data for each item of each assessment. In some cases, data were not used in this analysis. Reasons include repeated data and data that is missing but not at random. No missing data was considered missing at random as in most cases of missing data, a single item or multiple items were missing for a group of students in the same program and year. Data for this analysis was downloaded from Taskstream (2016) and organized in Excel (2016). In cases where completion status was uncertain, status was confirmed where possible using the school of education student database. Non-completers were removed from the analysis.

While there are six evidences used for teacher candidate score reporting, evidence one representing breadth of content and evidence 4 representing pedagogy knowledge and skills from a clinical perspective, were graded in such a way that is equivalent to pass/fail. Any unreconciled or not completed scores would be assumed to prevent completion of the program. In this case the expected variance of these scores is zero and there is nothing of interest to be analyzed for these two evidences because all relevant scores should look the same. For this reason, these two evidence scores were excluded from further analysis. From the remaining categories, the total number of items used to assess each candidate are listed below and measure each of the evidences, CDAP, and TGAP instruments respectively as shown in Table 3.

Table 3

Number of Items Measuring Each Category Along with a Description

Category	Description	# of items
Evidence 2	Depth of Content	8
Evidence 3	Ped. Knowledge and Skills: Planning	15
Evidence 5	Impact on Student Learning	20
Evidence 6	Leadership Advocacy and Professional Practice	8
CDAP	Candidate Dispositions Assessment Process	9
TGAP	The Teacher Growth and Assessment for Preservice	18
Total	Total items in Assessment system	78

Additionally, within the TGAP system for Pre-Service Profile, there are four constructs measured by the 18 criterions as shown in Table 4.

Table 4

Constructs within the TGAP

	# of items
Planning	3
Instruction	7
Assessment	3
Student Motivation and Management	5
Total	18

The constructs within the TGAP instrument for the purposes of this analysis were combined into the one TGAP factor as correlations between these items were very high. Additionally, a principle components analysis of only the TGAP items revealed that one factor explained almost 70% of the variance, with no other eigenvalues greater than 0.62 and when using an oblique rotation for correlated factors, the highest item loadings were

not always associated with the expected construct. The following table shows the number of responses in each category by item (Table 5):

Table 5

Response Counts by Item

	Score	Count	Percent
E21	2	563	67
	3	281	33
E22	2	580	69
	3	264	31
E23	2	570	68
	3	274	32
E24	2	578	68
	3	266	32
E25	2	568	67
	3	276	33
E26	2	554	66
	3	290	34
E27	2	534	63
	3	310	37
E28	2	535	63
	3	309	37
E31	2	459	54
	3	385	46
E32	2	460	55
	3	384	45
E33	2	507	60
	3	337	40
E34	2	487	58
	3	357	42
E35	2	462	55
	3	382	45
E36	2	447	53
	3	397	47
E37	2	531	63
	3	313	37
E38	2	508	60
	3	336	40
E39	2	509	60

	3	335	40
E310	2	439	52
	3	405	48
E311	2	506	60
	3	338	40
E312	2	506	60
	3	338	40
E313	2	569	67
	3	275	33
E314	2	493	58
	3	351	42
E315	2	469	56
	3	375	44
E51	2	531	63
	3	313	37
E52	2	586	69
	3	258	31
E53	2	589	70
	3	255	30
E54	2	541	64
	3	303	36
E55	2	541	64
	3	303	36
E56	2	533	63
	3	311	37
E57	2	591	70
	3	253	30
E58	2	532	63
	3	312	37
E59	2	559	66
	3	285	34
E510	2	545	65
	3	299	35
E511	2	578	68
	3	266	32
E512	2	517	61
	3	327	39
E513	2	558	66

	3	286	34
E514	2	574	68
	3	270	32
E515	2	524	62
	3	320	38
E516	2	552	65
	3	292	35
E517	2	670	79
	3	174	21
E518	2	589	70
	3	255	30
E519	2	520	62
	3	324	38
E520	2	498	59
	3	346	41
E61	2	503	60
	3	341	40
E62	2	542	64
	3	302	36
E63	2	449	53
	3	395	47
E64	2	550	65
	3	294	35
E65	2	520	62
	3	324	38
E66	2	465	55
	3	379	45
E67	2	512	61
	3	332	39
E68	2	477	57
	3	367	43
D1	3	6	1
	4	40	5
	5	213	25
	6	585	69
D2	3	16	2
	4	75	9
	5	231	27

	6	522	62
D3	3	10	1
	4	58	7
	5	222	26
	6	554	66
D4	3	8	1
	4	78	9
	5	280	33
	6	478	57
D5	3	9	1
	4	86	10
	5	232	27
	6	517	61
D6	3	12	1
	4	84	10
	5	214	25
	6	534	63
D7	3	10	1
	4	54	6
	5	214	25
	6	566	67
D8	3	14	2
	4	93	11
	5	279	33
	6	458	54
D9	3	9	1
	4	55	7
	5	218	26
	6	562	67
T1	3	23	3
	4	147	17
	5	295	35
	6	379	45
T2	3	10	1
	4	108	13
	5	278	33
	6	448	53
T3	3	10	1
	4	101	12

	5	278	33
	6	455	54
T4	3	18	2
	4	159	19
	5	285	34
	6	382	45
T5	3	18	2
	4	172	20
	5	296	35
	6	358	42
T6	3	19	2
	4	184	22
	5	298	35
	6	343	41
T7	3	17	2
	4	134	16
	5	312	37
	6	381	45
T8	3	19	2
	4	190	23
	5	286	34
	6	349	41
T9	3	22	3
	4	202	24
	5	303	36
	6	317	38
T10	3	21	2
	4	197	23
	5	305	36
	6	321	38
T11	3	23	3
	4	221	26
	5	284	34
	6	316	37
T12	3	17	2
	4	126	15
	5	282	33
	6	419	50
T13	3	21	2

	4	150	18
	5	302	36
	6	371	44
T14	3	25	3
	4	146	17
	5	297	35
	6	376	45
T15	3	15	2
	4	106	13
	5	268	32
	6	455	54
T16	3	19	2
	4	133	16
	5	287	34
	6	405	48
T17	3	16	2
	4	175	21
	5	311	37
	6	342	41
T18	3	16	2
	4	100	12
	5	290	34
	6	438	52

Note here that whereas a response of 2 was more common than a response of 3 on every item in the evidences, the CDAP and TGAP items are skewed towards the upper end with every item having the highest number of scores in the highest category. This pattern is especially prevalent in the CDAP items as the percentage of scores in the highest categories is higher in most cases. It is important to have a sufficient number of scores in each category included in the analysis in order to produce stable results. The natural implication of this skewness is that the CDAP and TGAP items are easier than the evidence items to achieve a higher score. One connection with a later result in this

analysis, is that item information is maximized at the difficulty thresholds of an item, thus a lower difficulty would have higher information nearer the lower end of the ability scale. The result of this difference in difficulty causes the CDAP and TGAP items to have higher information near the passing threshold than the evidences since the threshold is at the lower end of the ability scale in this case.

Preliminary Analysis

The preparation for the research questions began with some preliminary analysis of the data. This included checking the data for normality and multidimensionality, and checking specifically dissattenuated correlations between the factors, as well as reliability using Cronbach's alpha.

The data appear to violate multivariate normality. Even univariate normality, a prerequisite to multivariate normality, seems to be violated as measured by several items demonstrating high levels of skewness and kurtosis and P-P plots that show departures from normality in several places. The data show aspects of both unidimensionality (reliability almost 97%) and multidimensionality (as a principle components analysis revealed six factors with an eigenvalue over two and the first factor explaining about 30% of the data). The assessment system intends to measure overall teaching capacity, as well as various correlated factors that were believed to make up teaching capacity. To remove the error in the correlation coefficient that is due to the unreliability of the assessments, a disattenuated correlation coefficient was calculated between each pair of factors. Correlations were moderate between the evidences, weak between the evidences and the CDAP and TGAP respectively, and strong between the CDAP and the TGAP.

Reise, Moore, and Haviland (2010) discussed the phenomenon of assessments (especially those intended to measure different aspect of a single overlying trait) displaying characteristics of both unidimensionality and multidimensionality. It is suggested that bi-factor models are particularly useful "for evaluating the plausibility of subscales, determining the extent to which scores reflect a single variable even when the data are multidimensional, and evaluating the feasibility of applying a unidimensional IRT measurement model" (Reise et al., 2010, p. 1). It is suggested that in such contexts applying a unidimensional model to multivariate data can produce interpretable results and be appropriately fitted to a unidimensional model.

Data Preparation

In some instances, both scaling and the removal of repeated or missing data allowed for clearer interpretation of results. These types of changes can be summarized in four ways: re-scaling some data to its intended scale, eliminating duplicate data, rounding data in order to better represent the intended ordinal scale or the data and eliminating accidental non-completers.

In some programs the CDAP and TGAP scores were recorded on a 3-point scale where the instructions were to record evaluation scores on a 6-point scale. In order to compare all programs on the same scale these scores were doubled. Since scores were also not always recorded as integers but were either estimated as belonging between two rubric criteria or averaged from multiple evaluators, this doubling caused data to fall on the six-point scale. However, once modified, it did not always equal a 4 or a 6 precisely. In another case, scores were recorded on a 3-point scale where 0, 1 and 2 were used

instead of 1, 2 and 3. In this case, one was added to each score before scores we doubled. Additionally, in one program, the CDAP scores for 2011-12 and 2012-13 were recorded based on an older rubric. Scores were reconciled on items where the rubric was the same, but where item wording deviated, scores were treated as missing data.

It was discovered that data from the NC Teach program recorded identical data to those in the English program with the same student recorded twice with the same score. In these cases, a single copy of the scores was included and the duplicate data was removed from the final database.

Failing scores were recorded in the final data however they were not used in the analysis. In these cases, scores were sent back to the data source to be confirmed as to their status. If the failing score prevented completion, the data was eliminated from the analysis, since the population of interest only includes completers. If the failing score did not prevent completion, the score was rounded up to the lowest passing score. In order to preserve the intended ordinal scale, in cases where non-integer scores were recorded, final scores were rounded to the nearest available score indicated on the rubric, in this case, the nearest integer score. This was done in order to collapse small categories that contained low frequencies, and provide sufficient data to produce stable estimates of item thresholds.

Methods

This analysis begins by fitting the data to several CFA models. Although the data are ordinal and non-normal a robust weighted least squares method (WLSMV) method of estimation will be used. The WLSMV method is described as the best option for

modeling categorical or ordered data due in part to the fact that a normality assumption is not required (Brown, 2015; Proitsi et al., 2011). WLSMV is the default method used in Mplus for categorical data and Albright (2006) confirms this method of estimation as similar to the diagonal weighted least squares (DWLS) method implemented by Lisrel. Once the best fitting model was retained, item loadings and thresholds were interpreted. Then this model was used to evaluate two types of multigroup analysis that will explore first, factorial invariance between two types of evaluators, and then a MIMIC model will test the items for uniform DIF across levels of evaluators.

Assessing Model Fit

In nested models fit indices can be compared directly by interpreting an absolute difference in either values of certain fit indices, or a chi-square difference test of significance can be performed. This is roughly significant if the ratio of the difference in the chi-square values and the degrees of freedom in the model is greater than about three. This is also known as the likelihood-ratio (LR) test (Bollen, 2014). Cheung and Rensvold (2002) suggested assessing the change in the CFI (Bentler, 1990), gamma hat, and McDonald's (1989) Noncentrality index when comparing multiple groups for measurement invariance. Here he suggested that if the difference in the CFI was less than or equal to 0.01, the model did not fit significantly worse and should not be rejected. Little (1997) suggested four criteria when comparing the relative fit indexes of two nested models. They are goodness of overall fit, a max difference of 0.05 in the Tucker–Lewis Index (TLI), misfit is uniformly and unsystematically distributed with respect to

the constrained parameters, and the constrained model is both more meaningful and parsimonious than the unconstrained model.

Some indices are sensitive to various aspects of the model or the data, such as sample size and model complexity. For example, chi-square will often reject models with large sample sizes even when true differences are practically insignificant. Even difference tests of chi-square between models can be sensitive to sample size in a similar way. It is important to consider multiple indices in order to get a good sense of model fit when comparing models. Satorra and Bentler (2001) continued to work on improving modifications to the chi-square difference test in order to better accommodate conditions such as non-normal data. According to a study by Hutchinson and Olmos (1998), the indices that performed best under a variety of design conditions when used to look at ordered categorical data were the CFI, the incremental fit index and the non-normed fit index referred to as the TLI.

In non-nested models, comparison is not as straightforward. When two models are not nested a chi-square ratio test is no longer appropriate but rather indices such as the CFI, RMSEA and the BIC can be used to compare between models. Comparison of non-nested models is further complicated when the data is non-normal since several indices of fit can function differently when data is not normal. Clarke (2003) presented one example of research in comparing non-nested models does not assume normality. In cases of non-nested non-normal data, the overall fit of the models can be assessed using multiple indices, considering the most parsimonious model, when fit is similar.

Research Question 1

Using the results of preliminary analysis as a guide, the research begins with a comparison of model fit between a single factor model, a bi-factor model, a higher order model, a six-factor correlated model, and a six-factor orthogonal model where factors are not allowed to correlate. A one-factor model would represent unidimensional data, where all items loaded primarily onto one key factor represented by teaching capacity. If the model best fits the data as determined by multiple fit indices and as compared to the alternative models, it would reveal the contribution of each item to the general factor. The six-factor model alternatively would represent a multi-dimensional structure where six factors are being measured by this assessment system and the best way to interpret the contribution of an item would be to observe its contribution to the specific factor that it loads on such as teacher dispositions or teacher growth. A bi-factor model lets each item load both on a general factor and specific factors representing a scenario where one general factor is being measured but there is still covariance in the data beyond that explained by the general factor that is addressed by the specific factors. A six-factor correlated model acknowledges the correlations of the factors but represents the scenario where the factor correlations are not sufficiently explained by a single general factor, implying multiple explanations not specified in the model. Finally, a higher order model represents a scenario where each specific factor is believed to be an aspect of the general factor teaching capacity where the specific factors are thought of as separate measures of a single latent trait. Another way of looking at this model is to view the general factor as explaining all the correlations between the specific factors aside from error.

In each case fit will be assessed using multiple indices including the chi-square, root mean squared error of approximation (RMSEA), the CFI, TLI, and the weighted root mean-square residual (WRMR). The chi-square tests the discrepancy between the sample and fitted covariance matrices. The RMSEA tells us how well the model with unknown but optimal parameter estimates would fit the population covariance matrix. The CFI is a revision of the Normed-fit index and compares the covariance matrix with a null model in which case all latent variables are uncorrelated. The TLI is similar to the CFI; however, it increases the penalty for model complexity. Interpreting multiple fit indices helps account for multiple aspects of fit to the model and can help paint a clearer picture when answering precisely and justifiably one of the most important questions of structural equation modeling; does the model fit the data (Hooper, Coughlan, & Mullen, 2008)? Some suggested cutoff values to determine good fit for these indices include greater than or equal to 0.95 for the CFI and TLI, less than or equal to 0.06 for the RMSEA, and less than about 1.0 for the WRMR (Hu & Bentler, 1999; Newsom, 2012; Yu, 2002). The WRMR is considered appropriate for data that is non-normal and categorical such as in this model (Cook, Kallen, & Amtmann, 2009).

Research Question 2

The next research question considers what we can tell about the items and the instrument once a structure is imposed on the data. To explore this question, using the model from research question one, the slope parameters or loadings onto the general factor (or specific factors in the case of the six-factor model) can be interpreted as a measure of item quality. The intercepts or threshold parameters can be interpreted as a

type of item difficulty. Using these parameters, item and test information and characteristic curves can be created. The standardized item loadings on a factors show how well each item measures the intended factor. The item characteristic function and the item information function can help determine how well the items are distinguishing between people along the ability scale for each item. The item thresholds for each item represent the difficult of an item specifically in moving from one category to another. For example, when assessing categorical data, we can represent an item characteristic curve with multiple curves on a graph. These curves represent the probability (which sum to one at any ability) of a candidate having each categorical score depending on the latent ability trait. The item information function, on the other hand, will help us determine not only how well the items are distinguishing between people of various abilities, but also where on the ability scale (for a general or specific factor) each item is performing the job best.

Holt (2014) showed the relationship between the factor analysis threshold and loading parameters, with the IRT difficulty and discrimination parameter for both dichotomous and polytomous items using the conversion formulas

$$a_j = d(\frac{\lambda_j}{\sqrt{\psi_i}}) \text{ and } \quad b_{jc} = \tau_{jc} / \lambda_j \qquad (5)$$

where a_j is the discrimination parameter, λ_j is the loading onto the general factor, ψ_i is the error variance for item j and d is the scaling factor of 1.702 which converts the parameter from the normal ogive graded response model IRT parameter to the more

common logit scale of the logistic graded response model IRT parameter. Then b_{jc} is the item difficulty parameter for item j and category c, and τ_{jc} is the item threshold for item j and category c.

The item characteristic curves (ICC) and the IIC use the parameters of an item to demonstrate certain characteristics of that item across the ability scale. In the case of the ICC the probability of falling into a certain response category is shown across the ability scale, and in the case of the IIC, the amount of information or precision is shown across the ability scale. For the WLSMV method of estimation used by Mplus, the ICCs are represented by the following formulas:

if j is the first category

$$P_{ijk}(f) = P(U_i = j|f, G = k, X = x) = \Phi\left(\frac{\tau_{ijk} - \lambda_{ik}f - \beta_{ik}x}{\sqrt{\theta_{ik}}}\right) \qquad (6)$$

If j is the last category

$$P_{ijk}(f) = P(U_i = j|f, G = k, X = x) = 1 - \Phi\left(\frac{\tau_{ij-1k} - \lambda_{ik}f - \beta_{ik}x}{\sqrt{\theta_{ik}}}\right) \qquad (7)$$

If j is a middle category

$$P_{ijk}(f) = P(U_i = j|f, G = k, X = x) = \Phi\left(\frac{\tau_{ijk} - \lambda_{ik}f - \beta_{ik}x}{\sqrt{\theta_{ik}}}\right) - \Phi\left(\frac{\tau_{ij-1k} - \lambda_{ik}f - \beta_{ik}x}{\sqrt{\theta_{ik}}}\right) \qquad (8)$$

where G is the grouping variable, Φ is the standard normal cumulative distribution function, U_i is a categorical indicator for the latent factor f, and X represents other covariates or latent variables.

IC have the ability to tell us where across the ability scale the measurement error is lowest, in other words how precise is our measure for different levels of ability. IICs are calculated in Mplus (Muthén & Muthén, 2010) using the formula

$$I_{ik}(f) = 3.29 * \frac{\lambda_{ik}^2}{\theta_{ik}} \sum_{r=1}^{l} \frac{(Q_{irk}(1-Q_{irk}) - Q_{i,r-1,k}(1-Q_{i,r-1,k}))^2}{P_{irk}} \qquad (9)$$

for the WLSMV estimation method where P_{irk} represents the item characteristic curve function (Asparouhov & Muthén, 2015):

$$Q_{ijk} = \sum_{r=1}^{j} P_{irk} \qquad (10)$$

Research Question 3

In the next stage of this analysis a multi-group SEM analyses was performed using the TGAP and CDAP data where multiple evaluators are present. The model was the equivalent model from the developed model used in the full data set, which in this case is a two factor correlated model. This multi-group analysis seeks to determine whether differences exist between scores recorded by university supervisors and onsite teacher evaluators. Reliability was calculated for each factor for each group using ordinal alpha that was adjusted for ordinal data by substituting polychoric correlation coefficients

in the formula for Cronbach's alpha instead of Pearson correlation coefficients in the formula

$$\alpha = \frac{(k*r)}{(1+(k-1)*r)} \qquad (11)$$

where k is the number of items and r is the average polychoric correlation coefficient (Gadermann, Guhn, & Zumbo, 2012). Before differences in latent variables can be interpreted, invariance of the groups must be established on some level, in other words does the same model fit for both groups. If this is not the case, then apparent differences in latent means or variances are not necessarily due to true differences but could be a result of differences in the item parameters across groups. Invariance is often tested step by step at multiple levels beginning with configural invariance. In this test no parameters are constrained but freely estimated in each group to see if the same pattern of observed variables loading onto general or specific factors is observed in each group. Once this was established the next step was to constrain the factor loading and or the threshold parameters in each group to be equal. This is sometimes done together and sometimes done separately. Testing equality of factor loadings is often called metric invariance and testing of the item thresholds is referred to as a test of scalar invariance. If metric invariance holds, weak invariance is established and it can be appropriate to consider differences in the latent means. If both metric and scalar invariance hold, then strong invariance is established and differences observed in the latent mean can confidently be interpreted. One final test that was not considered necessary for interpretation of latent

differences was strict invariance which considers equality of residual variances (Byrne, 2008; Oliden, 2011).

A chi-square difference test is often used to determine whether each consecutive model fits significantly worse than the previous model, however recommendations of tests using other indices exist such as assessing the change in the CFI, gamma hat, and McDonald's Noncentrality index (Cheung, 2002). When comparing models used under the estimation method of WLSMV, or when distributional assumptions about the data are not met, chi-square cannot be compared with a simple difference of ratio test as is commonly implemented. Instead an adjustment must be made to the chi-square statistic to compare two models (Satorra & Bentler, 2010). The adjustment was made automatically using the Mplus software.

For the multigroup analysis, only data from the TGAP and CDAP instruments contain multiple evaluators. The model will test for structural, scalar, and metric invariance across groups and if this holds, differences in the latent means were examined and interpreted. The data for this analysis will consist of 560 students from several programs who were scored by both an OSTE and a university supervisor independently scored, with final scores representing averages. The results of this analysis will demonstrate any differences in the structural model for the two groups and potentially and differences in raters that exist for the CDAP and the TGAP instruments.

Research Question 4

Finally, using the developed structural model a MIMIC model will be used to detect potential uniform DIF. This analysis seeks to determine whether items can be

identified that may be interpreted differentially by the two groups of evaluators supervising elementary education candidates and middle/secondary candidates. These groups respectively will be treated as the focal and the reference groups. A group variable will be added to the final model, which will contain a zero for data points in the reference group (middle/secondary) and a one for data points in the focal group (elementary). These variables loadings were estimated freely onto each factor, then placed in the model as having loadings fixed to zero on each item. The modification index to the model will be observed noting any loading path with an index greater than 3.84. From here the item with the highest modification index will be freed to be estimated by the model and the process will be repeated until no modification indices are above 3.84. This method of DIF detection was similar to that used by Proitsi et al. (2011) when detecting the influence of multiple covariates on behavior and psychological symptoms in dementia. If there is cause to examine DIF for any items with respect to evaluators in elementary education programs and those in middle school/secondary education programs, the analysis should highlight those items. Flagged items (or lack thereof) will be noteworthy because they will identify items that may be interpreted differentially for the two groups.

The results of this method was compared to the results from Polysibtest under the condition where all items are suspected of DIF, in order to see how much overlap exist in items that are flagged. Under most circumstances in which a DIF detection method is used, there is reason to suspect DIF in certain items a priori. However, because this analysis was meant to serve more as an example rather than to confirm pre-specified

concerns about item DIF with this instrument, a correction to the acceptable p-value cut off must be implemented to account for testing 78 items in Polysibtest (Chang, Mazzeo, & Roussos, 1996). The correction used was the Bonferroni correction, which divides the critical alpha value by the number of items. For this analysis this adjusts the critical alpha value from 0.05 to about 0.00064 for the purposes of flagging items for DIF detection.

There exist other methods of detecting DIF in structural equation modeling, some of which detect both uniform and non-uniform DIF. The MIMIC model approach applied in this way only searches for uniform DIF which is consistent across the latent trait. Non-uniform DIF on the other hand would be present if the difference in the difficultly between the two groups was not consistent for all ability levels for a particular item across the two groups. In terms of an item characteristic curve, no item DIF means that the item characteristic curve is the same for both groups. Uniform DIF is represented by the groups having two characteristic curves of the same shape, but one shifted to the left or right across the ability scale. The amount of shift can be interpreted as the amount of DIF. Non-uniform DIF is present if the ICC for each of the two groups are of different shapes meaning that the amount of DIF and even the direction is dependent on the latent ability (Osterlind & Everson, 2009).

The results of these analyses provide evidence for the retained factor structure of the data, the quality of the items and any potential differences between raters. Also any items showing the potential existence of uniform DIF between elementary and middle/secondary evaluators will be identified.

CHAPTER IV

RESULTS AND DISCUSSION

A Comparison of the Proposed CFA Models

The first question posed by this research was the following: which CFA model can be used to appropriately represent the structure of this assessment system given the data? Three non-nested models, a bifactor, a higher-order model, and a correlated six-factor mode were tested and compared looking at differences in overall fit. Then the retained model was tested against the nested models including a six-factor orthogonal model and a single factor model. Each CFA was performed with categorical data using the estimation method WLSMV in Mplus and fit indices were compared as appropriate. The results of the five CFA models are shown below (Table 6 and Table 7), including an adjusted chi-square difference test for the nested models as compared to the six-factor correlated model, which was retained as well as standards for interpreting the fit indices used.

Table 6

A Summary of Fit Indices for the Five Tested Models

	χ^2	df	Difference Test (X2)	df change	RMSEA	CFI	TLI	WRMR
One Factor	34539.07	2925	2402.4	15 ($p \leq 0.01$)	.11	.79	.79	7.96
Six factor	24662.21	2925	1619	15 ($p \leq 0.01$)	.09	.86	.86	7.84
Bifactor	12460.22	2847			.06	.94	.93	3.70
Higher-order	12473.29	2919			.06	.94	.94	3.74
Six factor correlated	4137.49	2910			.02	.99	.99	1.07

Table 7

A Summary of Appropriate Cut Offs for Fit Indices

	RMSEA	CFI	TLI	WRMR
Poor Fit	>.1			
Acceptable fit	<.08	>.9	>.9	
Good Fit	<.05	>.95	>.95	<1

Adapted with permission from Browne, 1993; Newsom, 2012; Yu, 2002

Since the one-factor and six-factor model were nested in the other three models (which are not nested within each other), it makes sense to view this comparison in two stages. First, the non-nested models were compared consisting of the bi-factor model, the higher order model and the correlated six-factor model. Then, the chosen model from this first step was compared with the two nested models. When comparing the six-factor

correlated model with the bi-factor model and the higher order model, the fit of the six-factor model is better across all parameters. In order to compare them we can look at quality of model fit in each case, and in close cases accept the model that is most parsimonious. Out of these three models, the most parsimonious is the higher order factor model. However, the higher order model does not exhibit the best fit. The bifactor model and the higher order factor model have very similar fit, generally in the range of overall acceptable fit for the RMSEA, CFI, and TLI. However, neither of these models exhibited good fit according to the WRMR. On the other hand, the six-factor correlated model showed excellent fit for the RMSEA, CFI, and TLI and is just shy of good fit for the WRMR. A correlated bifactor model was tested and initially did not converge. However, when the convergence criteria were adjusted, the model fit slightly better than the six-factor correlated model. When the model was analyzed, however, all the correlations between the specific factors became significant, many of the loadings of the items onto the general factor were not significant, and some were negative and significant making the interpretation of the loadings on the general factor questionable. Sawaki, Stricker and Oranje (2009) found a similar result when comparing a correlated bi-factor model to a correlated traits model and concluded that this kind of result indicates a problem with model identification and the model was not developed further.

With the six-factor correlated model as our initially retained model, we can now compare this model directly with the one factor and the six-factor orthogonal models, since these models were nested within the six-factor correlated model. The six-factor orthogonal model is simply a more restrictive version of the six-factor correlated model

where all correlations were fixed to zero. The one-factor model is similar in that it is the same model as the six-factor correlated model where all factor correlations were fixed to one. Because these models were nested we tested differences using a modified chi-square difference test in addition to comparing the other fit indices. Even though the differences in chi-square can be inflated, as well as the chi-square statistic itself for large sample sizes such as this, it is important to note that the chi-square difference test alone may not be sufficient evidence to draw conclusion of practical differences between the models. Another test suggested by Cheung (2002) recommended considering one model fitting significantly worse than the other did if the difference in the CFI is greater than 0.01. In this case, when the fit indices of the six-factor correlated model are compared with the one factor and the six-factor orthogonal model, the chi-square difference is significant, and the CFI difference is greater than 0.01. While the six-factor correlated model demonstrates good overall fit, none of the indices show good or even moderate fit for either the one factor or the orthogonal six-factor model. Given these results, the six-factor correlated model will be developed, interpreted and discussed further regarding its use in modeling the student teacher candidate assessment data at UNCG.

Interpreting the Six-factor Correlated Model

Having established the six-factor correlated model, the next step is to interpret the parameters in order to gain insight about the items and the assessment. Research question two asks the following: when a structure is imposed on the model, what can we tell from the results about the items and the assessment? With the six-factor correlated model retained, there are now four things of interest to highlight here; the strength of the

factor loadings, the correlations between the factors, the thresholds, and the reliability of each factor. The following table (Table 8) shows the standardized loadings of the items onto each of the six factors. High loadings can be interpreted as representing well the intended factor and are a measure of item quality. All loadings are significant at (p<0.001).

Table 8

The Standardized Item Loadings onto the Six Factors

	E2	E3	E5	E6	D	T
E21	.90					
E22	.92					
E23	.90					
E24	.89					
E25	.93					
E26	.95					
E27	.93					
E28	.93					
E31		.94				
E32		.94				
E33		.95				
E34		.95				
E35		.96				
E36		.95				
E37		.96				
E38		.89				
E39		.94				
E310		.95				
E311		.95				
E312		.97				
E313		.98				
E314		.95				
E315		.95				
E51			.90			
E52			.85			
E53			.92			
E54			.90			
E55			.93			
E56			.95			
E57			.92			
E58			.92			
E59			.94			
E510			.96			
E511			.97			
E512			.94			

E513	.97			
E514	.97			
E515	.95			
E516	.96			
E517	.91			
E518	.95			
E519	.89			
E520	.94			
E61		.93		
E62		.95		
E63		.93		
E64		.86		
E65		.92		
E66		.91		
E67		.94		
E68		.95		
D1			.90	
D2			.87	
D3			.91	
D4			.85	
D5			.87	
D6			.91	
D7			.90	
D8			.93	
D9			.90	
T1				.89
T2				.91
T3				.84
T4				.90
T5				.87
T6				.89
T7				.83
T8				.95
T9				.89
T10				.90
T11				.90
T12				.93
T13				.89
T14				.89
T15				.92

T16	.90
T17	.85
T18	.88

The high loadings for all items is evidence which supports the current method of score reporting, which is reporting a profile score representing the six factors rather than a single score that represents teaching capacity. All of the loadings are above 0.8 and nearly all of them are above 0.85. Each item does a good job of representing the factor that it intends to measure and is of good quality with respect to the purpose of the instrument as measuring six factors.

The factor correlations are presented in Table 9. The story is very similar to that of the correlations corrected for attenuation presented earlier. The CDAP and TGAP factors are highly correlated, while most other factors show low to moderate correlations. Specifically, moderate correlations are shown between evidence factors, and weak correlations are shown between the CDAP and the evidences and between the TGAP and the evidences. The lowest correlations are shown between Evidence 2 and the TGAP, while the correlation between Evidence 2 and the CDAP is also weak. The evidences measure specifically the results of a project or paper graded by a supervisor, where the CDAP and TGAP are measure of performance and characteristics demonstrated by action taken in the classroom setting. For this reason, it is not surprising that the correlations between the evidences and the CDAP and TGAP are not strong, as they tend to measure a somewhat distinct set of skills.

Table 9

Factor Correlations

	E2	E3	E5	E6	D
E3	.43				
E5	.54	.51			
E6	.47	.45	.57		
D	.10	.26	.17	.21	
T	.06	.25	.17	.22	.81

The thresholds can be interpreted as the amount of the latent trait interpreted as a *z*-score (which in this model is the ability of the candidate with respect to the specific factor corresponding with the item) which is required to move from one specific score category to the next. As the latent trait increases, the probability of falling into the various score categories changes. For the evidences, there are only two categories, and for the CDAP and TGAP there are four, so each item has a number of thresholds equal to the number of categories minus one. Since there are six latent abilities measured by this assessment, it only makes sense to interpret an item threshold in terms of the specific latent ability corresponding to that item. To illustrate this, item E22 has a threshold of 0.488 and item E31 has a threshold of 0.11. However, these thresholds should not be compared directly because they represent a measure based on different abilities, where E22 at 0.488 and E27 at 0.339 can be compared since both are measured based on the same latent trait. The item threshold summary statistics are listed below:

Table 10

Item Threshold Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation
E2	8	.15	.34	.49	.42	.06
E3	15	.4	.05	.45	.20	.11
E5	20	.59	.23	.82	.41	.13
E6	8	.31	.08	.39	.24	.11
D1	9	.38	-2.45	-2.08	-2.26	.11
D2	9	.46	-1.6	-1.14	-1.33	.15
D3	9	.4	-.51	-.11	-.33	.13
T1	18	.38	-2.26	-1.89	-2.03	.10
T2	18	.56	-1.12	-.56	-.84	.17
T3	18	.42	-.1	.32	.12	.14

The threshold parameters can be thought of as a measure of difficulty where lower values mean easier items when compared within the same latent ability trait. The cut-score for a test that is in any way viewed as pass/fail, is an important concept to discuss when addressing the idea of precision across the ability scale. Since the thresholds will represent the points across the ability scale where the precision of the test is highest, it is desirable that the location of those thresholds be close to the location of the cut-score on the ability scale. Although only completers are included in the data set, the cut-score would represent the lowest score that could be achieved by a completer. Therefore, the lower the threshold on this scale, the closer to the cut-score it would be. This would not be the case if data were present which represented candidate scores below the cut-score. In such a case the exact point could be estimated more precisely. The lowest threshold for the CDAP and TGAP in this case come closest to the cut-score, where the threshold value for each of the evidences would be considered far from the cut-

score and therefore less useful in determining precise ability of a candidate whose true ability was near the cut-score.

As demonstrated before, the reliability coefficients for the six factors ranged from 0.925 to 0.975, which shows very high reliability within each factor. As another measure of reliability, a factor information curve demonstrates the level or levels of ability where each assessment does the best job of distinguishing an accurate measure of teacher candidate's latent ability. In other words, high information areas show where the standard error of estimation is smallest. The factor information curves are shown below followed by the information curves for the items making up each factor (Figures 1-12):

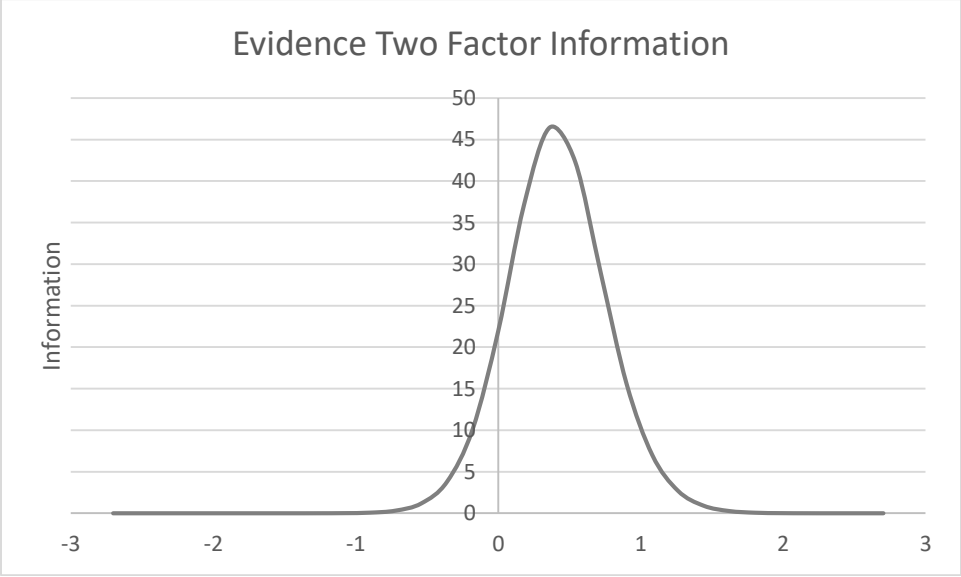

Figure 1. Evidence Two Factor Information Curve.

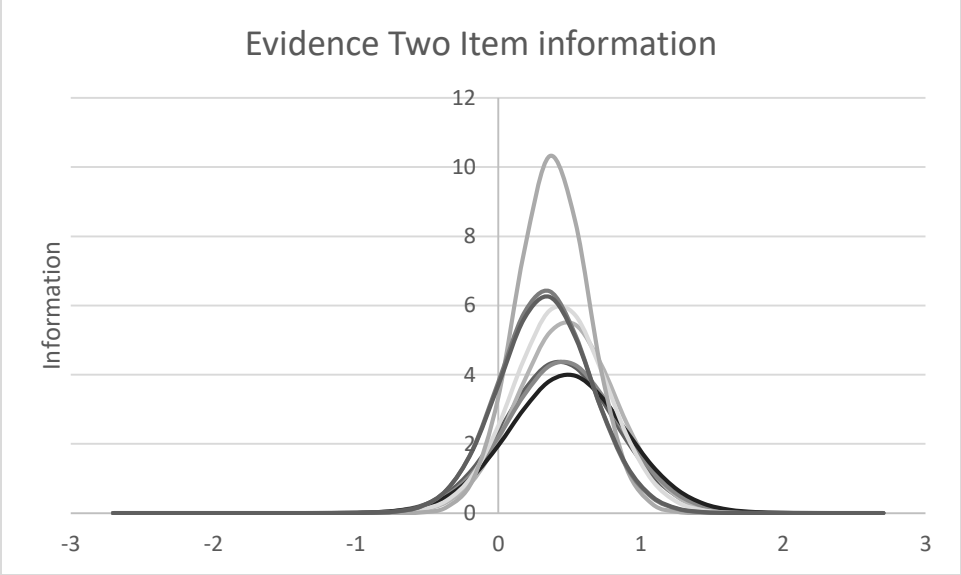

Figure 2. Evidence Two Item Information Curves.

Figure 2 shows the item information functions for the eight items represented by the Evidence Two factor.

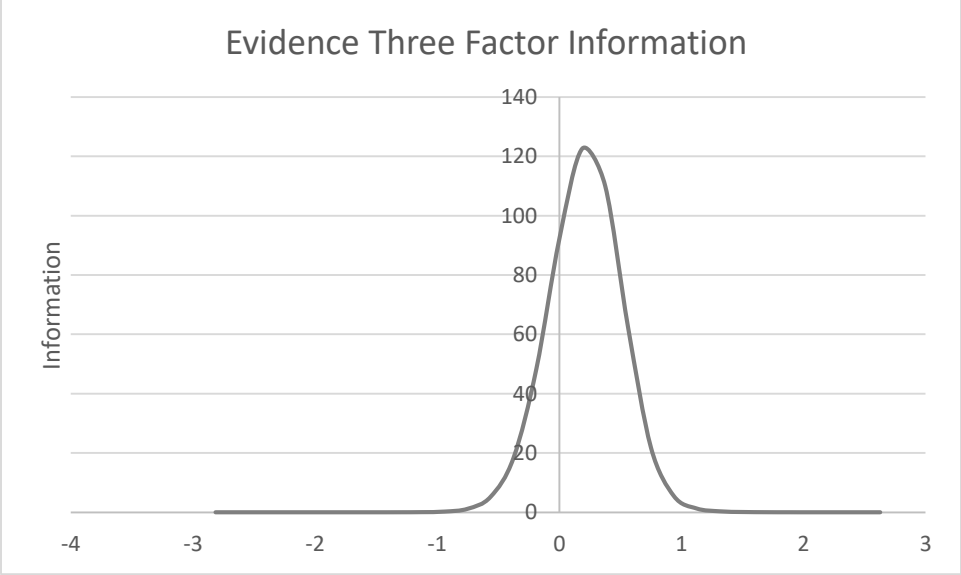

Figure 3. Evidence Three Factor Information Curve.

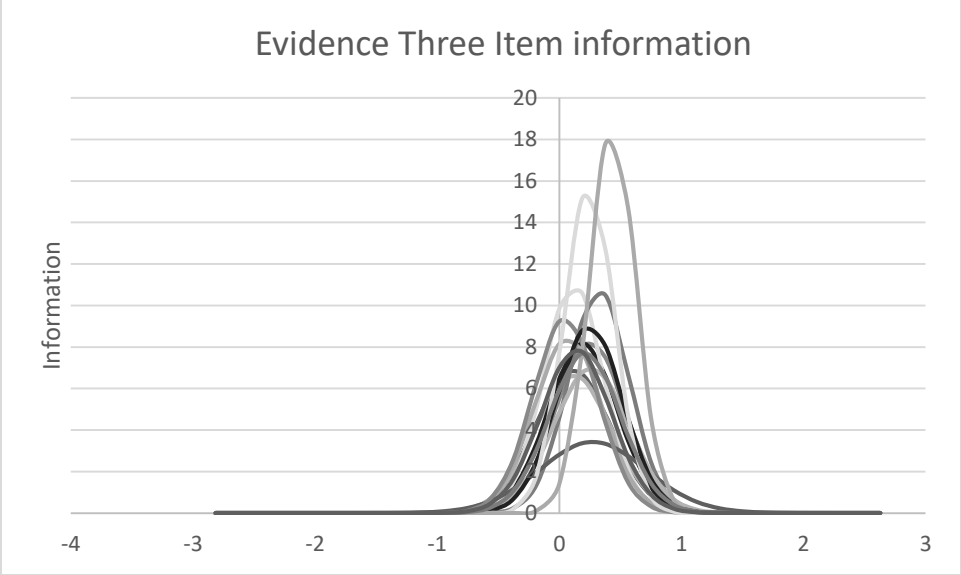

Figure 4. Evidence Three Item Information Curves.

Figure 4 shows the item information functions for the fifteen items represented by the Evidence Three factor.

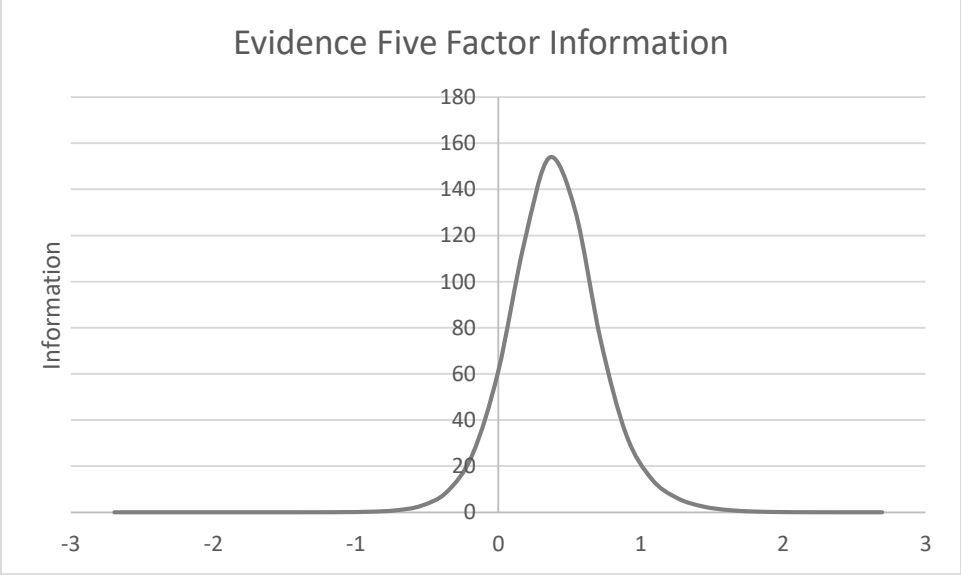

Figure 5. Evidence Five Factor Information Curve.

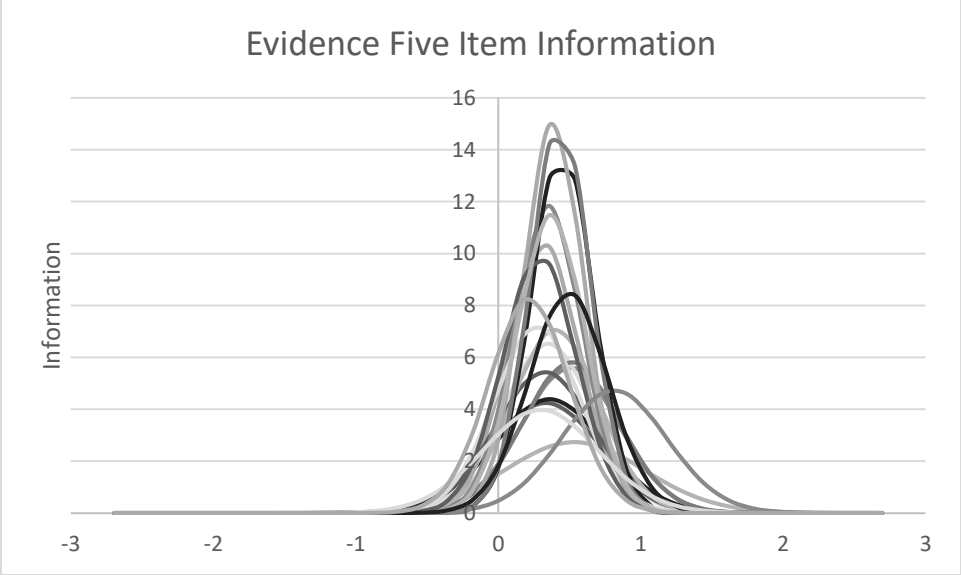

Figure 6. Evidence Five Item Information Curves.

Figure 6 shows the item information functions for the twenty items represented by the Evidence Five factor.

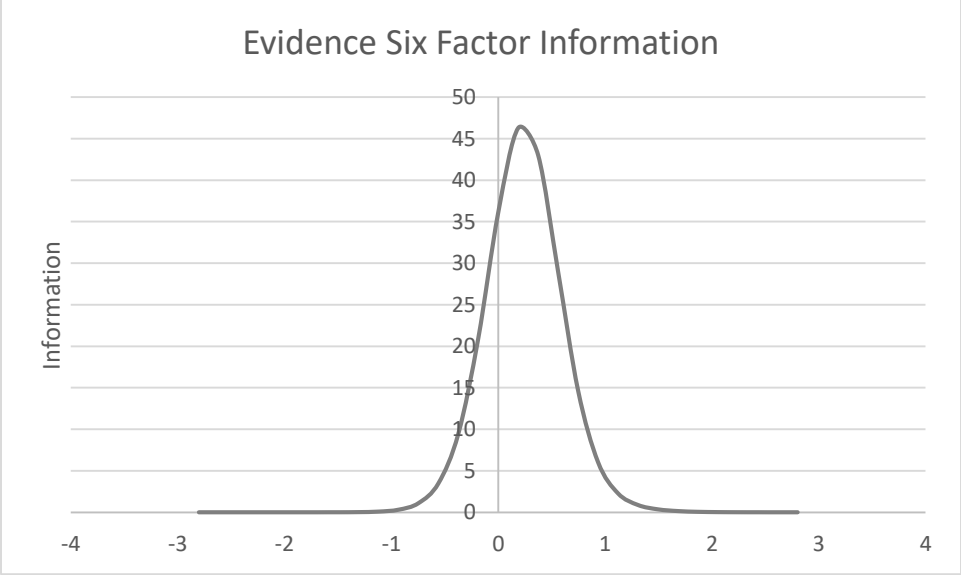

Figure 7. Evidence Six Factor Information Curve.

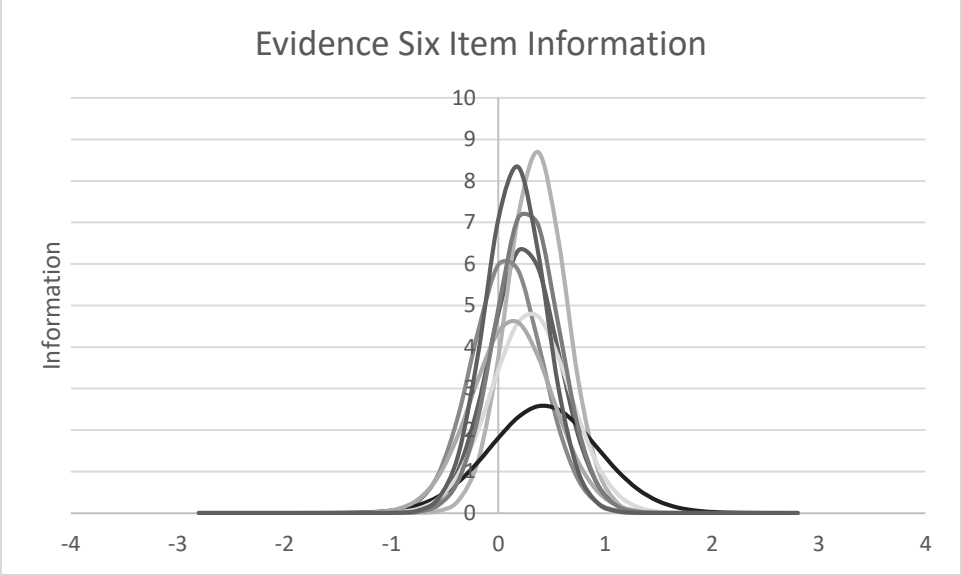

Figure 8. Evidence Six Item Information Curves.

Figure 8 shows the item information functions for the eight items represented by the Evidence Six factor.

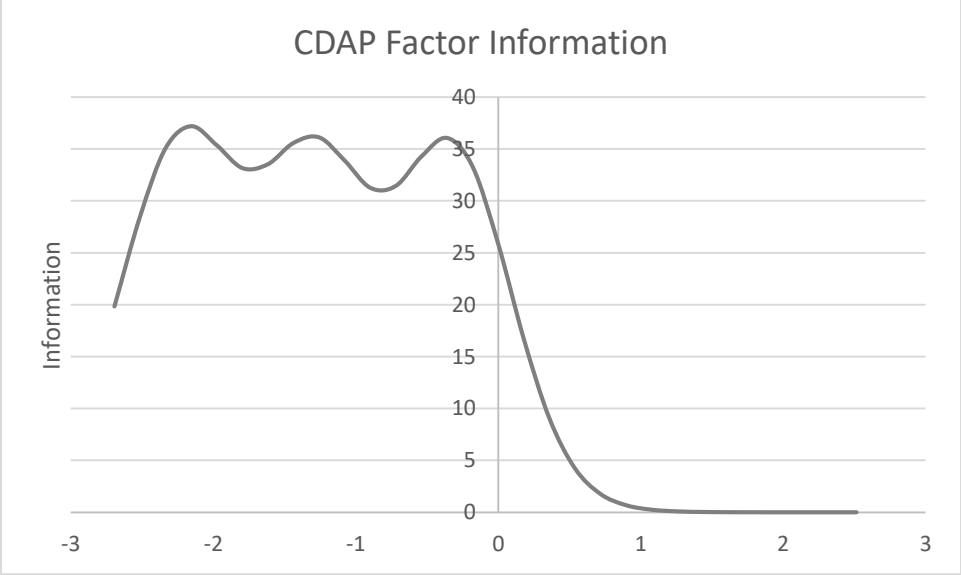

Figure 9. CDAP Factor Information Curve.

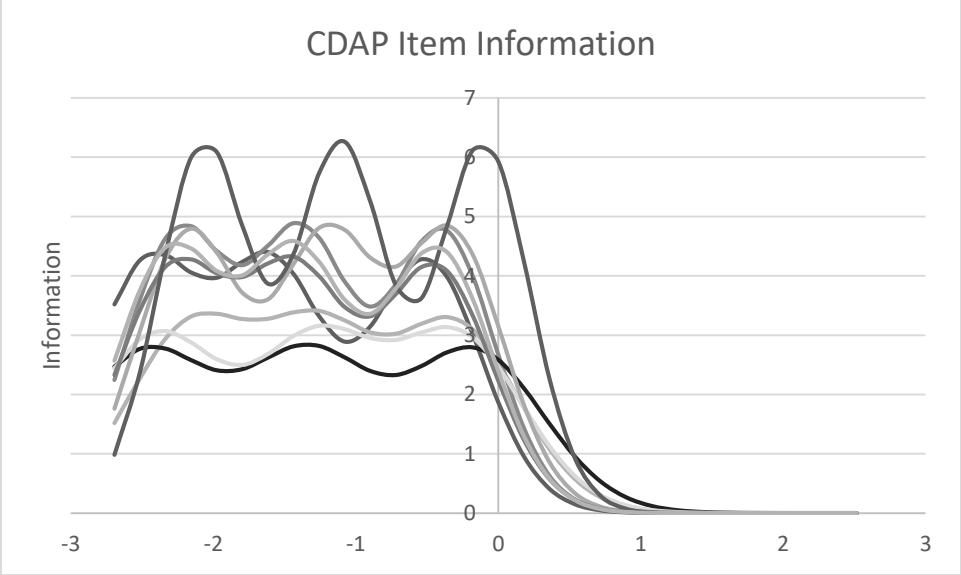

Figure 10. CDAP Item Information Curves.

Figure 10 shows the item information functions for the nine items represented by the CDAP factor.

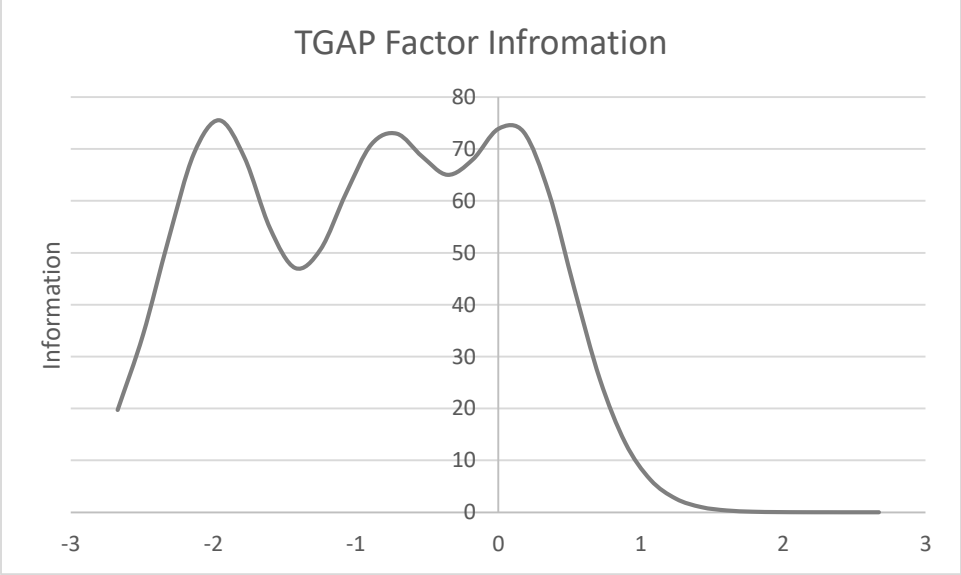

Figure 11. TGAP Factor Information Curve.

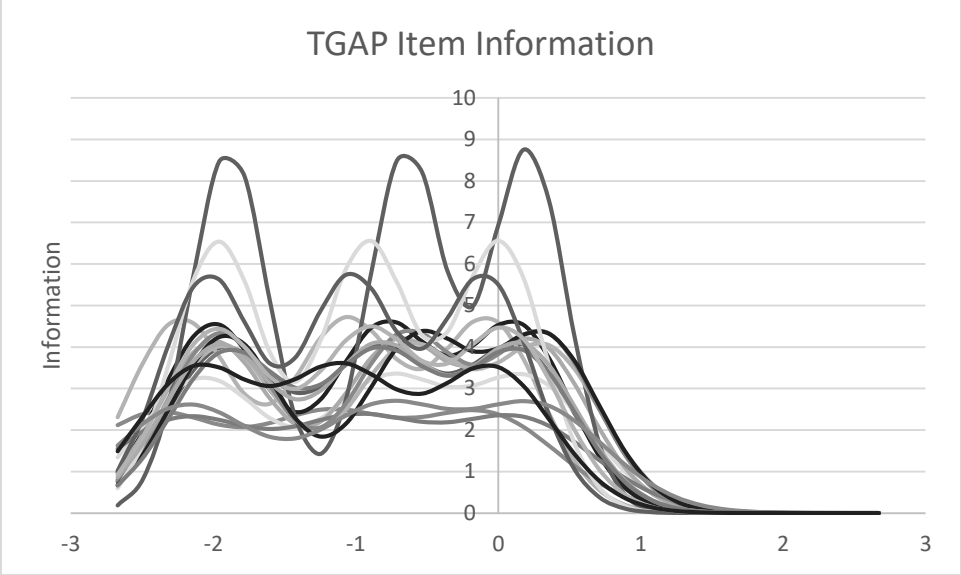

Figure 12. TGAP Item Information Curves.

Figure 12 shows the item information functions for the eighteen items represented by the TGAP factor.

Because the disposition and TGAP instruments have multiple score categories, the information curve peaks once for each threshold estimate. Where a cut-score is present, it is important to have high information near the cut score in order to take advantage of this low standard error. This can be summarized with the idea that if someone of very high ability is asked an easy question, or if someone of very low ability is asked a difficult question, very little about their ability level is learned that is not already know. In such a case, barring exceptional circumstance, the outcome is predictable. The only way to pinpoint an ability level is to ask a question where the difficulty is close to a candidate's true ability level.

In this case the cut-score could be represented by the lowest possible score for a completer, which would be a score of two for all evidence items and a score of three for all CDAP and TGAP items. Since any value below these cut-scores would result in failure, and this data contains only completers, the exact cut-score can only be estimated as being at the far left portion of the included charts. In line with this estimate, one important note here is that information for the CDAP factor is much higher near this cut-score than information near the cut-score for the evidences, where the TGAP could be considered somewhere in between but peaking nearer to the cut-score. The information for all the evidences peaks just above the average ability level of completers and is very low across the ability spectrum elsewhere. The implication of this is that the evidences are currently of little use in pinpointing ability near the cut-score, and the TGAP and CDAP provide some information nearer this cut-score but still do not peak at this point where precision is most important.

The Inner-rater Reliabilities of the CDAP and TGAP

Next we determined whether differences exist in the scores given by the supervising teacher and the cooperating teacher and if so, what is the nature of this difference. Using a corresponding model to the six-factor correlated model, the next step was to test for invariance of the model over multiple groups and ultimately test if applicable, the difference in latent means. Multiple rater data were only present for the items making up the CDAP and TGAP factors and only in certain programs. The six-factor correlated model fits the entire data, so the model that will be used for the multi-groups analysis will be a correlated two factor model.

When comparing two different groups for invariance, fit should be assessed for each group separately in order to make sure that the model itself makes sense to represent each group appropriately. A series of tests can then be conducted imposing more and more restrictions to the model across groups that tests invariance at each level. If at any point, throughout the process invariance does not hold, the process ends. Then based on the highest level of invariance that holds, interpretations of latent means or variances are considered. If the basic structure holds for each group, that is the items tend to load on the same factors and the same number of factors exist, also called configural invariance, then the next step is to see if the factor loadings themselves are equal across groups. This can be tested by imposing a constraint of equality across groups of the factor loadings and then see if the model fits significantly worse than the model where the loadings are freed across groups. If this model does not fit significantly worse than the less constrained model, then it is considered reasonable to assume that differences in the latent variables are not due to differences in the loadings. Additionally, the researcher can conclude that weak invariance, also called metric invariance, holds.

The next step in order to establish strong or scalar invariance is to constrain the threshold parameter to be equal across groups and test in a similar way whether the model fits significantly worse than the previously established model. If scalar invariance holds, then differences observed in the latent means are not due to difference in the thresholds across groups, but can be interpreted as true differences between the groups. One more test of invariance can be performed but is not considered necessary for interpreting the latent mean differences, which is called strict invariance. This was tested

by restricting the model further to constrain the residual variances of the observed variables to be equal across groups. Invariance was tested using the same indices used to establish good fit; however, often in this case, differences are interpreted. Cheung (2002) suggested that while a common test used to establish invariance is the chi-square ratio test, when sample size is large it may be more appropriate to consider differences in other indices to establish invariance because in the case of a large sample size the chi-square may show statistically significant differences where practical differences do not exist. In such a case, the difference in CFI can be used where a drop of less than or equal to 0.01 in the more restrictive model should be interpreted as an indicator of invariance. Table 11 shows the results of the tests for invariance performed where one groups consists of the scores given by the teacher candidate's supervising teacher and the other group consists of scores recorded by the on-site teacher evaluator. The chi-square differences were adjusted for the use of the estimator WLSMV.

Table 11

Results of the Tests for Invariance Across Evaluator Groups

	X2	df	Ratio Test	RMSEA	CFI	TLI	WRMR
Sup Group	1182.24	323		.07	.99	.99	1.17
OSTE	785.78	323		.05	.99	.99	.92
Configural	1970.73	646		.06	.99	.99	1.49
Metric	2016.31	671	52.33; *df=25*; p<.01	.06	.99	.99	1.50
Scalar	2044.97	723	93.75; *df=52*; p<.01	.06	.99	.99	1.54

When separated, both groups have good fit to the two factor correlated model. By the RMSEA and the WRMR however, the groups slip below the excellent fit mark. Scalar invariance seems to hold as demonstrated by overall good fit by all indices except for the WRMR and the chi-square. When testing for metric and scalar invariance the chi-square differences are both statistically significant. The CFI and the TLI are lowered by the more restrictive models by less than or equal to 0.01 in both cases, and the WRMR fit gets a little worse and the RMSEA fit gets a little bit better. In factoring all of these indices and using Cheung's (2002) criteria for the CFI difference no greater than 0.01, both the weak invariance and strong invariance models are retained as not fitting worse than the configural model. The differences were interpreted in the latent means as true differences in the latent means, and not differences caused by differences in the thresholds attributed to the observed variables. The reliability was high for both factors

in both groups. This reliability was calculated using ordinal alpha which is similar to Cronbach's alpha however instead of using the Pearson correlations, this formula inserts in its place the polychoric correlation matrix accounting for the ordinal nature of the data. Two factors may contribute to the magnitude of this reliability estimate. First, for the CDAP and TGAP, the high end of the scale is often used to score students; in some cases over 90% of the scores for an item are scored as a 5 or a 6, additionally in every case the scores are skewed to the higher end, with many items containing more than half the observed scores in the category of six. Additionally, when ordinal data is treated as continuous and estimated using Cronbach's alpha especially in the cases of skewed items, it is common for reliability to be underestimated. Using ordinal alpha can often produce higher estimates of reliability in such a case as this one (Gadermann et al., 2012) (Table 12).

Table 12

The Reliability of Each Group for Each Factor Calculated Using Ordinal Alpha

	CDAP	TGAP
Supervisor	.99	.99
OSTE	.98	.99

To look at differences in latent means, the standardized solution fixes the value of the supervisor group to zero and then estimates the OSTE group with a significance test to see if latent mean scores are different.

Table 13

Difference in Latent Means from the Scalar Model

	Estimate	SE	Est/SE	Two tailed P-value
D	.12	.27	.46	.65
T	.38	.15	2.57	.01

The results (Table 13) show that while there is not a significant difference in the scores given by supervising teachers and on-site teacher evaluators for disposition scores, the teacher growth score was significantly different and the on-site teachers award higher scores to teacher candidates than supervising teachers.

Testing for DIF with a MIMIC Model and POLYSIBTEST

For the final analysis, a MIMIC was used to address the question: Are the difficulties of some items dependent on whether the teacher candidate is in an elementary education program or in a middle grade/secondary program (Finch, 2005)? This is one way of thinking of DIF and this model tests for DIF by adding to the original SEM model a group indicator variable. For this analysis, data from elementary education programs was designated as one group and middle/secondary programs were designated another group. In order to draw conclusions about whether or not group membership affects a particular item, we partial out any effects on the latent variables by allowing each latent variable to load onto the grouping variable. Once this effect was accounted for, we can see if any item loading on the group variable are significant. This was done by fixing each item loading onto the grouping variable to zero, then checking the modification index in the Mplus output to see whether estimating a specific loading would improve

model fit past a particular threshold. In this case, an item was flagged if the modification index was greater than 3.84. If any items were flagged, then the model was estimated again with only the loading freed to be estimated with the highest modification index and the rest fixed once again to zero. This process was continued until no item loading had a modification index greater than 3.84. This method for DIF detection was compared to results acquired using polysibtest where all items were suspected of DIF and the significant *p*-value was adjusted for 78 items to a threshold of 0.00064. The final items that were flagged from each analysis are recorded below along with their corresponding estimates (Table 14).

Table 14

Items Flagged for Potential Uniform DIF by Each Detection Method

DIF items	Flagged by	Mplus Estimate	SE	Polysibtest estimate Beta	SE
E51	Both	-.37	.06	-.14	.03
E52	Both	-.36	.07	-.14	.03
T18	Both	.28	.05	.30	.05
E517	Both	.30	.07	.11	.03
E24	Both	.22	.06	.12	.03
E22	Mplus	-.31	.07		
E21	Mplus	-.30	.07		
E35	P-Sibtest			-.13	.03
E315	P-Sibtest			-.13	.03
T14	P-Sibtest			.19	.05
T16	P-Sibtest			.22	.05

Positive Favors Elementary Group Negative Favors Sec/Middle Group

In total, seven items were flagged by the MIMIC model and nine items were flagged by Polysibtest, with five items overlapping. For each item that overlaps, the direction of the DIF is consistent, with items E51 and E52 suspected of favoring the secondary/middle group, and items T18, E517 and E24 suspected of favoring the reference group. From those items that were flagged for suspected DIF on only one method, items E21, E22, E35 and E315 are suspected of DIF favoring the secondary/middle group and T14 and T16 are suspected of DIF favoring the elementary group (Table 15).

Table 15

A Description of the Items Flagged for DIF by One or More Detection Methods

Scale/DIF items	Favors	Flagged by	Description
E5/E51	Sec/Middle	Both	Impact on Student Learning: Overview is clear and makes connections to the specific characteristics of the school that supports making instructional decisions for all students
E5/E52	Sec/Middle	Both	Impact on Student Learning: Assessment data related to multiple characteristics of student are analyzed and explained
TGAP/T18	Elem	Both	Student Motivation and Management: classroom climate
E5/E517	Elem	Both	Impact on Student Learning: instructional adaptations reflect collaboration with specialists
E2/E24	Elem	Both	In-depth Inquiry Project: Literature is rich, current, cited and relevant
E2/E22	Sec/Middle	Mplus	In-depth Inquiry Project: Significance of topic is explained
E2/E21	Sec/Middle	Mplus	In-depth Inquiry Project: Identified and articulates a topic in his or her academic field
E3/E35	Sec/Middle	P-Sibtest	Ped. Knowledge and Skills: Planning: Opportunities are provided to students to expose them to and help them understand multiple points of view
E3/E315	Sec/Middle	P-Sibtest	Ped. Knowledge and Skills: Planning: The product is presented clearly, is comprehensive and is well organized
TGAP/T14	Elem	P-Sibtest	Student Motivation and Management: Expectations/procedures
TGAP/T16	Elem	P-Sibtest	Student Motivation and Management: student interest and participation

More detailed statistics from the PolySibtest DIF detection method are provided below, the reference and focal columns represent the percentage of cases not used from each group in the analysis and the MS SSD was the matching subtest standardized score

difference, and the zero for each item in the flag column means that for every item a normal successful completion of a Sibtest run was performed (Table 16).

Table 16

Detailed Output of the PolySibtest DIF Detection Analysis

Item	Mean	Biserial	Beta	SE	*P*-value		Reference	Focal	MSSSD	Flag
E21	2.33	.34	-.09	.04	.01	E	.15	.17	.14	0
E22	2.31	.36	-.10	.04	.00	E	.16	.18	.14	0
E23	2.33	.34	.02	.03	.56	E	.18	.19	.14	0
E24	2.32	.38	.12	.03	0	E	.15	.19	.14	0
E25	2.33	.33	-.02	.03	.49	E	.14	.16	.14	0
E26	2.34	.37	.02	.03	.51	E	.14	.16	.14	0
E27	2.37	.36	.06	.03	.06	E	.15	.15	.14	0
E28	2.37	.38	.05	.03	.12	E	.13	.13	.14	0
E31	2.46	.53	-.11	.04	.00	E	.16	.17	.14	0
E32	2.46	.50	-.04	.03	.20	E	.15	.16	.14	0
E33	2.40	.52	-.01	.03	.84	E	.15	.14	.14	0
E34	2.42	.51	-.07	.03	.03	E	.15	.16	.14	0
E35	2.45	.54	-.13	.03	0	E	.16	.17	.14	0
E36	2.47	.50	-.07	.04	.06	E	.17	.16	.14	0
E37	2.37	.52	-.01	.03	.86	E	.15	.14	.14	0
E38	2.40	.48	-.04	.03	.18	E	.16	.18	.14	0
E39	2.40	.56	-.09	.03	.00	E	.15	.16	.14	0
E310	2.48	.49	-.06	.03	.06	E	.16	.16	.14	0
E311	2.40	.52	.03	.03	.34	E	.18	.17	.14	0
E312	2.40	.53	-.10	.03	.00	E	.18	.15	.14	0
E313	2.33	.55	.02	.03	.53	E	.17	.17	.14	0
E314	2.42	.53	-.08	.03	.02	E	.15	.14	.14	0
E315	2.44	.53	-.13	.03	0	E	.18	.18	.14	0
E51	2.37	.48	-.14	.03	0	E	.13	.15	.14	0
E52	2.31	.44	-.14	.03	0	E	.16	.15	.14	0
E53	2.30	.50	-.06	.03	.03	E	.13	.14	.14	0
E54	2.36	.48	-.06	.03	.07	E	.13	.16	.14	0
E55	2.36	.50	-.05	.03	.12	E	.16	.18	.14	0
E56	2.37	.51	-.03	.03	.42	E	.14	.15	.14	0
E57	2.30	.49	-.02	.03	.45	E	.15	.15	.14	0

E58	2.37	.50	-.10	.03	.00	E	.16	.17	.14	0
E59	2.34	.52	-.03	.03	.29	E	.15	.17	.14	0
E510	2.36	.55	.07	.03	.02	E	.11	.16	.14	0
E511	2.32	.55	.07	.03	.02	E	.12	.15	.14	0
E512	2.39	.52	.02	.03	.51	E	.15	.16	.14	0
E513	2.34	.54	.03	.03	.26	E	.12	.15	.14	0
E514	2.32	.54	.05	.03	.08	E	.12	.15	.14	0
E515	2.38	.53	.01	.03	.81	E	.13	.15	.14	0
E516	2.35	.54	.01	.03	.66	E	.15	.16	.14	0
E517	2.21	.40	.11	.03	0	E	.13	.15	.13	0
E518	2.30	.52	.02	.03	.46	E	.14	.15	.14	0
E519	2.38	.53	-.01	.03	.75	E	.14	.16	.14	0
E520	2.41	.55	-.04	.03	.22	E	.16	.16	.14	0
E61	2.40	.43	-.10	.04	.01	E	.15	.16	.14	0
E62	2.36	.45	.01	.03	.74	E	.11	.14	.14	0
E63	2.47	.45	.02	.04	.60	E	.15	.18	.14	0
E64	2.35	.41	.07	.03	.03	E	.16	.18	.14	0
E65	2.38	.44	-.08	.04	.02	E	.15	.16	.14	0
E66	2.45	.45	-.08	.04	.04	E	.13	.18	.14	0
E67	2.39	.46	.06	.03	.10	E	.16	.16	.14	0
E68	2.44	.45	-.04	.04	.22	E	.12	.14	.14	0
D1	5.63	.56	-.05	.04	.24	E	.17	.17	.14	0
D2	5.49	.59	.00	.05	.93	E	.22	.16	.14	0
D3	5.56	.56	-.04	.05	.39	E	.17	.17	.14	0
D4	5.46	.56	-.09	.05	.05	E	.19	.16	.14	0
D5	5.49	.57	.03	.05	.60	E	.18	.17	.14	0
D6	5.51	.62	-.02	.05	.63	E	.2	.2	.14	0
D7	5.58	.57	.01	.04	.78	E	.19	.17	.14	0
D8	5.40	.62	.04	.05	.48	E	.17	.18	.14	0
D9	5.58	.58	-.02	.04	.71	E	.18	.17	.14	0
T1	5.22	.66	.14	.05	.01	E	.16	.18	.13	0
T2	5.38	.64	.15	.05	.00	E	.19	.22	.13	0
T3	5.40	.60	-.02	.05	.73	E	.12	.16	.14	0
T4	5.22	.65	.03	.06	.55	E	.22	.2	.14	0
T5	5.18	.62	-.01	.05	.81	E	.15	.18	.14	0
T6	5.14	.64	.06	.06	.26	E	.17	.17	.13	0
T7	5.25	.64	-.04	.05	.43	E	.14	.16	.14	0
T8	5.14	.71	.13	.05	.01	E	.21	.2	.13	0
T9	5.08	.66	.16	.05	.00	E	.17	.2	.13	0
T10	5.10	.65	.13	.05	.01	E	.13	.17	.13	0

T11	5.06	.69	.07	.05	.16	E	.14	.18	.14	0
T12	5.31	.67	.10	.05	.03	E	.16	.17	.13	0
T13	5.21	.65	.05	.05	.37	E	.18	.18	.14	0
T14	5.21	.62	.19	.05	0	E	.19	.21	.13	0
T15	5.38	.65	.14	.05	.00	E	.14	.18	.13	0
T16	5.28	.65	.22	.05	0	E	.12	.17	.13	0
T17	5.16	.61	.10	.05	.06	E	.12	.17	.14	0
T18	5.36	.63	.30	.05	0	E	.18	.16	.13	0

Discussion

Contradictory to the hypothesis that the bifactor model would be the model that fit the UNCG teacher candidate assessment data, the six-factor correlated model had the best fit considering multiple fit indices and parsimony. When interpreting the six-factor model, it is important to consider what is meant by each model that was tested and how these interpretation relates to how the instrument is viewed. Structural equation modeling, or specifically CFA, gives us the ability to statistically test the different ways that we might view how the instrument is structured to see which model the data seems to support.

For each model tested, there was a different interpretation of the assessment system and using SEM the interpretations were justified by confirming that the data were well represented. The interpretation of the one-factor model in light of this assessment system would be that there is one common construct being measured by all 78 items and one score could adequately summarize the performance of a teacher candidate, an ability which might be called teaching capacity. A six-factor uncorrelated model could be interpreted as the test system measuring six important, unique and different constructs, each of which is measured by a specific set of items. In such a case a candidate's scores

are best represented by a profile, that highlights their ability on each of the factors separately, because it would make sense for a candidate to potentially be strong in one of these areas and weak in another.

Each of the last three models assumes that there exists some relationship between the measured factors but they were interpreted differently. When factors are related in an instrument, it is possible for the instrument to measure both one overall factor and several unique factors. For example, one might consider geometry and algebra. The basic rules of math apply to both and thus having a high level of competence in arithmetic would help a candidate on both tests; however, there is also a unique element to each concept that does not overlap with the other and requires a more specific skill. For this reason, one could be talented in one, both, or neither, but a weakness in the basics of mathematics such as division and multiplication would hinder someone from success in both areas. A six-factor correlated model allows for the factors to correlate and gives no single explanation, which would account for why some factors correlate and others don't, implying that multiple explanations may exist within these correlations. A bi-factor model attempts to explain the commonality in the factors with one general factor, then allows the remaining portion of the unexplained variance to be accounted for with specific factors. If a bi-factor model fits well it is not common for factors to be correlated beyond that which is explained by the general factor. If that were the case, then there is some other explanation beyond ability in the general factor (in this case teaching capacity), which causes correlation between the specific factors; however, the uniqueness of the factors above and beyond teaching capacity in general is preserved by the bi-factor

model. On the other hand, the higher order model constrains the factor correlation structure to assume a common underlying dimension across the six factors. In this case, the commonality of the factors in the model was explained entirely by the general factor and high loadings of the specific factors onto the general factor imply that the specific factors do not reflect abilities beyond the general factor.

The six-factor correlated model was retained as the best fitting model to the data, and one way to further develop this model is to consider why some factors may have correlated more strongly or more weakly with each other. The four evidences all had moderate correlations within themselves and represented scores graded by a university supervisor that was based on a rubric of a specific project completed and submitted by the teacher candidate. The highest correlation between factors was between the CDAP and TGAP factors. This is not surprising since these factors represent in several cases the average of two and in rare cases, more than two evaluators. Also the CDAP and TGAP are both assessed multiple times during a teacher candidate's preparation experience, where the evidence portfolios are evaluated only once. The lowest correlation between factors exists between evidence two and the TGAP. This is an interesting note conceptually as Evidence Two was the depth of knowledge project where a teacher candidate demonstrates deep knowledge of a subject area with emphasize on all the details of an academic paper, where the TGAP is measured based on the perception of both the supervising teacher and the OSTE of the teacher candidate's performance and growth over the course of their student teaching and preparation experience. One might view this as a difference in academic knowledge and experiential knowledge.

The fit of the correlated six-factor model in general is evidence for the reporting of a profile score as opposed to a single score and the high loadings of all items onto their corresponding factor is good evidence for the quality of these items in representing these factors. When a profile score is reported, it can easily highlight candidate scores that were not consistently high or consistently low across factors, revealing specific areas for improvement or relative weakness for a teacher candidate. This has the advantage of highlighting a candidate's strengths and weaknesses more specifically than a summary score, but it also can highlight potential areas within the programs where candidates are meeting the requirements of specific factors with a varying level of proficiency, increasing the specificity of feedback for both the candidate and the programs. Although the candidate is ultimately given a pass or fail in regard to licensure recommendation, this profile reporting would not likely impact a candidate's potential for employment. One interesting observation about the factor information curves was that all the evidences seem to have their peak a little above the average ability, where information for the CDAP and TGAP factors have higher information below the average ability. Because all scores represent completers, failing scores are not present in the data. Since it is necessary to pass every item in order to complete the program, if enough data were collected from those that did not pass to have sufficient data in the failing categories to estimate a model, the IIC could help us discern the precision of each item near the cut score.

When the scores were analyzed across the different rater groups it was found that the factor structure holds for both the supervising teachers and the on-site teacher

evaluators, as well as strong invariance. Furthermore, TGAP scores were higher in the OSTE group as compared to the supervising teacher scores, but no statistically significant difference was found in scores across the CDAP factor. This result is evidence for inner rater reliability across the disposition factor but not the teacher growth factor. Porter and Jelinek (2011) found inter-rater reliability measures for the performance assessment for California teachers to be moderate to poor, but it worsened overall in candidates who failed the assessment as compared to those that passed. If scores were compared for candidates who failed, it would be interesting to see if this trend was consistent with Porter's results.

The MIMIC model found seven items flagged as suspected of having DIF, five of which overlapped with the Polysibtest results when the focus and reference group were defined as those scored by elementary supervisors and those scored by middle grade/secondary supervisors. Normally, DIF detection is used as a confirmatory analysis where there is reason to believe that some items have DIF where others do not. When DIF detection is used to explore all items on an assessment, flagged items should be considered and evaluated further before conclusions about DIF are drawn. Two possible suggestions for this process might include re-reading carefully the items flagged and consider whether there is good reason based on the wording of the item to suspect the item for DIF. If the problem seems obvious, corrections to the item wording could be made. Alternatively, when additional cohorts are exposed to this item, further testing could be done with a new sample, and these results could be used to help determine which items might be suspected of DIF a priori. It is important to acknowledge that the

DIF found in this example is not necessarily cause for alarm, as some differences in the instructional methods and evaluation of teacher candidates are likely to occur between elementary instructors and middle secondary instructors, especially since the number of items favoring each group is similar in that all the items are not stacked against one particular group. This result, however, is interesting and should open the door for further exploration of the items in context and the rubric, that is used equally in both groups.

Although these findings provide evidence for the quality of the assessment system and the items, there are several highlights that can open discussion for improvement on the assessment system itself. Also, many more questions are raised, so this research sets the stage for researchers to tackle and further develop. The last chapter addresses a few of these questions that could be answered with preparation and specific data collection, to provide schools of educations with the even more of the tools they need to answer an evolving set of questions and requirements, and continue to ensure the best preparation experience possible for our teacher candidates.

CHAPTER V

SUMMARY, FUTURE RESEARCH, AND CONCLUSIONS

Summary of Results

As demonstrated by multiple indices of fit, the six-factor correlated model, best represents the factor structure of the teacher candidate assessment system. When this factor structure is used to model the data, the loadings of all items are very high, which is evidence that each item does a good job of measuring the construct it is intended to measure. The evidence instruments, as shown by the information curves do the best job of accurately measuring completer abilities just about the average ability level, and the CDAP and the TGAP instruments do the best job of distinguishing completer ability below the average ability level. Inter-rater reliability was shown to be good between the OSTE evaluators and the supervising teacher evaluators for the dispositions however, the OSTE evaluators gave better evaluation scores in the area of teacher growth than the supervising teachers. Finally, the MIMIC model, demonstrated that some of the items flagged for the possibility of DIF, overlaps with the results from PolySibtest, although both methods flagged some items that the other method did not. Overall this analysis provides evidence for the reliability of the instruments used, support for the scoring profile, and evidence for the quality of the items. These finding invite further exploration into the reasons for the possible presence of uniform DIF in a few items.

While the six-factor correlated model fit the data best as opposed to the expected model which was the bi-factor model, evidence was found to support the quality of the instrument overall. The purpose of this project can be summarized with a three-step process. The first step establishes the structure of the data. The second step uses that structure to assess the quality and attributes of the items and the assessment system as a whole. The third stage, with the structure established, develops that structure to answers questions concerning group differences. Both the multi-group test for invariance and differences in latent means, and the MIMIC model used to detect DIF are examples of the more advanced statistical methods that can be used to evaluate teacher candidate assessments. The results between the multi-group model and the MIMIC model are not directly comparable here, since the groups were divided differently, but it could be interesting to compare the two methods with the same division of groups in future research.

Implications

The implications of this research can be divided into two categories. First the implications for the school of education at UNCG and second for the field of education. The impact this research has on the school of education at UNCG is that it provides evidence for the quality of the items in measuring their intended factors and also provides feedback in that item information is generally not near the cut score especially for the evidences. This could be addressed by experimenting with expanding the scale of the evidences to a six-point scale rather than a three-point scale and including scores from candidates who did not pass every item and were not recommended for licensure. The

implications of this research for the field of education add a fresh application of statistical techniques to teacher candidate assessment including the use of SEM, polysibtest, and MIMIC modeling to address the evaluation of teacher candidates. With the addition of new statistical tools for use by teacher candidate evaluators, the most efficient, descriptive and precise answer to complex questions can be more adequately explored.

Limitations to the Study

Approximately 17.5% of the original data contained missing values, often with a distinct but inconsistent pattern. In some cases, a few single variables were missing. In other cases, a collection of questions representing an entire factor were missing. The way this was accounted for in the research was to eliminate all candidate data that had missing data, since the data was neither normal, nor did it appear to be missing at random. These holes in the data limit the generalizability to the programs as a whole considering in many cases programs which already contained small amounts of data were reduced further in sample size, increasing the percentage of data drawn from the larger programs. Additionally, because the data were not missing at random but, in fact, was often common to a few specific small program, this has potential to limit the generalizability of the results to include these specific underrepresented programs because the already small sample size of these programs within the data was reduced further by the presence of missing data. Furthermore, if a pattern existed within the data that was present for these programs with significant missing data, these possibly interesting results will be undetectable by this research due to the missing data. Scores were present in the data that were not appropriate for a completer to have obtained, for example, a score of one on any

evidence or a score of one or two on the CDAP or TGAP factors should have at least temporarily prevented completion. Even though it is possible these problems were addressed privately and resolved before completion, evidence of this was not present in the final reported database. Where failing scores were present, an additional database were used to confirm, where possible, completion status. Teacher candidates who did not complete were removed from the analysis. Since the category created by failing scores was so small, in order to produce stable results as well as preserve the intention of the instrument with regard to completers, all remaining failing scores for confirmed completers were rounded up to the lowest possible passing score. Although these instances were few, if these score were never reconciled before a candidate completed, rounding causes a loss in true variability of the data. Several non-integer scores were also present in the data. This makes sense when scores were averaged across multiple evaluators however, this only explained a portion of the non-integer scores. The presence of non-integer score violates the ordinal nature of the intended scale as measured by the rubric. The implication of non-integer scores is that the scale was interpreted as continuous by at least some evaluator while the rubric specifies an ordinal scale. This is potentially problematic since the type of estimator used in SEM depends on the nature of the data. For this analysis the rubric was chosen as the standard and all non-integer scores were rounded to the nearest integer score which was represented on the rubric. Additionally, it is important to keep in mind that due to the size of the programs in the teacher education program at UNCG, the programs were not equally represented, and no adjustment was made to attempt equal representation. For example, the elementary and

elementary GC programs combined represent approximately a third of the total teacher candidates used in this study. Program data was analyzed as a group for two reasons: first, because some programs had very few graduates in the span of 3 years and, secondly, because the rubric was not program specific. In other words, for the purposes of this assessment system, the rubric, scale, and criteria for each score category was the same for every program, implying that the intention was for the assessment to measure equally all programs.

While an adjustment was made to correct for type 1 error in the SIBTEST results, no correction was made in the MIMIC model to adjust for familywise type 1 error rate. Also the SIBTEST results did not account for the six-factor structure of the model while the MIMIC model did. Even though several fit indices were used to get a clear picture of overall model fit, one aspect of fit that could be addressed in future research would be to explore item and person fit. This could allow consideration of whether patterns exist in person misfit as well as the overall percent of person misfit. Also, this could highlight any items that stand out as not fitting well within the final model.

Future Research

This research scratches the surface of the applications of using advanced statistical methods to demonstrate assessment quality. As new training methods for evaluators are developed and requirements change, one direction of future research might explore the effect training has on evaluator scores and inner rater reliability. Different methods of training such as face-to-face and online training exists and, if enough data is present, feedback on the effects of different training styles, content, or methods, perhaps

incorporating interactions of these effects across different types of evaluators, such as gender, age, OSTE vs supervisor, etc. This could provide insight that could improve evaluator training in a way that provides the most benefit to a variety of evaluators. Another direction could explore further the assumption that the assessment system performs equally across programs, ultimately to justify or advocate against the use of a single rubric across programs. Another extension to this study could develop the evaluation of the validity of this assessment system. First, validity could be examined as represented by relationship of success on this instrument with success in the teaching profession. Depending on what data was available to the researcher, this could be defined through self-reported success as collected through surveys, evaluations of supervisors of the new teacher, average scores on end-of-course tests of students in the classroom of the new teacher, or simply persistence in the teaching field as measured by employment status over several years. If the last definition were adopted, new teachers would need to be tracked outside of North Carolina in addition to within, since there is currently no way to tell the difference between a teacher candidate who left the profession and one who simply left the state. Another way to explore validity would be to implement qualitative methods to explore the details of the assessment system itself, including the creation of the instrument and its implementation, results, and purposes. Surveys could be used in addition to interviews of employees, supervisors, teachers, college faculty, and even teacher candidates in order to highlight the experience of the teacher preparation program at UNCG.

One major strength of IRT based information curves is their ability to discern where on the ability scale, the items and the assessment as a whole are most reliable, and have the most accuracy in pinpointing specific ability levels. This is especially true in instruments where cut scores exist, and where high stakes decisions such as that of licensure recommendations are based on that cut score. In this way, information curves can justify the precision of instruments or items near the cut score, allowing additional confidence in the accuracy of these decisions. If data are maintained, which included teacher candidate score who did not pass, generalizations could be made about all teacher candidates who were prepared by the program, where the current research was limited to completers. With candidate data that include valid score categories both below and above the cut score, the IIC could observe the precision of the items near the cut scores, capitalizing on the advantages of IRT.

Recommendations

While the results presented by this research are specific to the assessment system at UNCG, the methodology and their applications are meant to inspire other universities to consider new ways of addressing questions both old and new within assessment systems, developing new tools to provide evidence of reliability, validity, and quality.

When establishing a scale for an assessment, it is less important what scale is used but rather that the scale that is used be interpreted consistently. In training, it is worth noting what is meant by a score of two or a three but also when none of the options given by the rubric fit the candidate exactly; it is important that the judgments made about scoring are consistent. For example, even if the definition of a score of two or three is

universally understood, what happens to the students who clearly belong somewhere between satisfactory and exceptional? Some evaluators may consider anything above a two to be rounded up to a three, others might consider anything below a three to be rounded down to a two, while still others may choose to simply implement their own scale and against the recommendations of the instructions, to record a value of 2.75. It is not in such cases so important which way values are rounded or how these problems are dealt with, but the problem arises if solutions to these problems are inconsistent across raters or programs especially since the number of categories is so low. Inconsistent rounding can have a large impact on results. Simply addressing this issue during training session can help improve consistency and thus the ability to interpret scores in the same way that are recorded by different evaluators or programs.

Measuring or demonstrating a relationship between success on the evaluation instrument and success in the teaching profession would be an excellent opportunity to show criterion validity. There are many ways to view success in the profession and any overlap with the data currently available should be considered with regard to addressing this relationship. Alternatively, new collection of data should be considered for the purposes of demonstrating this aspect of validity.

Conclusion

This research demonstrates a variety of applications of structural equation modeling to educational assessment, and while the results in many ways provide evidence for the overall quality of the teacher candidate assessment system at the UNCG, they also highlight several areas for discussion and future research to consider. Structural

equation modeling is one of many tools that can aid when used properly in continuously striving for improvement, and in both maintaining and exceeding a high standard for teacher candidates, and programs that prepare them to serve our children.

REFERENCES

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, 25(1), 95-135. http://dx.doi.org/10.1086/508733

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19,* 716–723. doi: 10.1109/TAC.1974.1100705

Albright, J. J. (2006). *Confirmatory factor analysis using AMOS, LISREL, and MPLUS.* Retrieved from http://www.iub.edu/~ statmath/stat/all/cfa/cfa2008. pdf

Almerico, G., Johnston, P., Henriott, D., & Shapiro, M. (2011). Dispositions assessment in teacher education: Developing an assessment instrument for the college classroom and the field. *Research in Higher Education Journal, 11*, 1. doi:10.1177/0022487109348024

Angus, D. L. (2001). *Professionalism and the public good: A brief history of teacher certification*. Retrieved from https://edex.s3-us-west-2.amazonaws.com/publication/pdfs/angus_7.pdf

Asparouhov, T., & Muthén, B. (2015). *IRT in Mplus.* Retrieved from https://www.statmodel.com/download/MplusIRT.pdf

Assessment & Support Consortium. (2011). *InTASC model core teaching standards: A resource for state dialogue*. Retrieved from intasc_model_core_teaching_standards_2011.pdf

Bacci, S., Bartolucci, F., & Gnaldi, M. (2014). A class of multidimensional latent class IRT models for ordinal polytomous item responses. *Communications in Statistics-Theory and Methods, 43*(4), 787-800. http://dx.doi.org/10.1080/03610926.2013.827718

Bales, B. L. (2006). Teacher education policies in the United States: The accountability shift since 1980. *Teaching and Teacher Education, 22*(4), 395–407. doi: dx.doi.org/10.1016/j.tate.2005.11.009

Barrett, J. (1986). *Evaluation of student teachers*. Washington D.C.: ERIC Clearinghouse.

Bates, A. J., & Burbank, M. D. (2008). Effective student teacher supervision in the era of No Child Left Behind. *Professional Educator, 32*(2), 1. doi:10.1257/jep.24.3.133

Benjamin, W. J. (2002). *Development and validation of student teaching performance assessment based on Danielson's framework for teaching*. Retrieved from http://files.eric.ed.gov/fulltext/ED471552.pdf

Bentler, P. M. (1986). Structural modeling and Psychometrika: A historical perspective on growth and achievements. *Psychometrika, 51*(1), 35-51. doi:10.1007/BF02293997

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238. doi: 10.1007/BF02293997

Bidwell, Allie. (2015, March 23). *Report: States need to connect teacher evaluations to other quality measures.* Retrieved from http://www.usnews.com/news/articles/2013/10/30/reportstatesneedtoconnectteacherevaluationstootherqualitymeasures.

Borko, H., & Mayfield, V. (1995). The roles of the cooperating teacher and university supervisor in learning to teach. *Teaching and Teacher Education, 11*(5), 501-518. Doi: 10.1016/0742-051X(95)00008-8

Bollen, K. A. (2014). *Structural equations with latent variables.* New York: John Wiley & Sons.

Bolt, D. M. (2000). A SIBTEST approach to testing DIF hypotheses using experimentally designed test items. *Journal of Educational Measurement, 37*(4), 307-327. doi:10.1111/j.1745-3984.2000.tb01089

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. New York: Guilford Publications.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. *Sage Focus Editions, 154*, 136-136. doi: 10.1177/0049124192021002005

Byrne, B. M. (2008). Testing for multigroup equivalence of a measuring instrument: A walk through the process. *Psicothema, 20*(4), 872-882. doi:10.1037//1099-9809.7

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment* (Vol. 17). Washington D.C.: Sage Publications.

Carnegie Forum on Education, & the Economy. Task Force on Teaching as a Profession. (1986). *A nation prepared: Teachers for the 21st century: The report of the task force on teaching as a profession.* New York: Carnegie.

Chang, Y. W., Huang, W. K., & Tsai, R. C. (2015). DIF detection using multiple-group categorical CFA with minimum free baseline approach. *Journal of Educational Measurement*, *52*(2), 181-199. doi: 10.1111/jedm.12073

Chang, H. H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement*, *33*(3), 333-353. doi:10.1111/j.1745-3984.1996.tb00496

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*,*9*(2), 233-255. doi: 10.1207/S15328007SEM0902_5

Clarke, K. A. (2003). A simple distribution-free test for nonnested hypotheses. New York: University of Rochester.

Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: A practical guide to Kane's framework. *Medical Education*, *49*(6), 560-575. doi: 10.1111/medu.12678

Cook, K. F., Kallen, M. A., & Amtmann, D. (2009). Having a fit: Impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. *Quality of Life Research*, *18*(4), 447-460. doi: 10.1007/s11136-009-9464-4

Council for the Accreditation of Educator Preparation. (2015, July 31). *History of CAEP*. Retrieved from http://caepnet.org/about/history

Council of Chief State School Officers. (2015, August 9). *Interstate teacher assessment and support consortium (InTASC).* Retrieved from http://www.ccsso.org/resources/programs/interstate_teacher_assessment_consortium_(intasc).html

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334. doi:10.1007/BF02310555

Damon, W. (2007). Dispositions and teacher assessment the need for a more rigorous definition. *Journal of Teacher Education*, *58*(5), 365-369. doi: 10.1177/0022487107308732

Danielson, C. (2011). *Enhancing professional practice: A framework for teaching*. Alexandria, VA: ASCD.

Danielson, C. (2013). *The framework for teaching*. Retrieved from https://www.danielsongroup.org/framework/

Darling-Hammond, L. (2000). Teacher quality and student achievement. *Education policy analysis archives*, *8*, 1. http://dx.doi.org/10.14507/epaa.v8n1.2000

Darling-Hammond, L., & Bransford, J. (2007). *Preparing teachers for a changing world: What teachers should learn and be able to do*. Hoboken, NJ: John Wiley & Sons.

Deng, L., & Yuan, K. H. (2015). Multiple-group analysis for structural equation modeling with dependent samples. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(4), 552-567. http://dx.doi.org/10.14507/epaa.v8n1.2000

Fang, Z. (1996). A review of research on teacher beliefs and practices. *Educational Research*, *38*(1), 47-65. http://dx.doi.org/10.1080/0013188960380104

Farkas, S., Johnson, J., & Duffett, A. (1997). *Different drummers. How teachers of teachers view public education. A report*. New York: Public Agenda.

Farrell, T. S. (2008). Here's the book, go teach the class' ELT practicum support. *RELC Journal*, *39*(2), 226-241. doi: 10.1177/0033688208092186

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement*, *29*(4), 278-295. doi: 10.1177/0146621605275728

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*(4), 466. doi: 10.1037/1082-989X.9.4.466

Fukuhara, H., & Kamata, A. (2011). A bifactor multidimensional item response theory model for differential item functioning analysis on testlet-based items. *Applied Psychological Measurement*, *35*(8), 604-622. doi: 10.1177/0146621611428447

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research & Evaluation*, *17*(3), 1-13. http://pareonline.net/

Goldhaber, D., Liddle, S., & Theobald, R. (2013). The gateway to the profession: Assessing teacher preparation programs based on student achievement. *Economics of Education Review*, *34*, 29-44. http://dx.doi.org/10.1016/j.econedurev.2013.01.011

Harper, C. A. (1939). *A century of public teacher education: The story of the state teachers' colleges as they evolved from the normal schools*. New York: American Association of Teachers Colleges.

Henry, G. T., Thompson, C. L., Bastian, K. C., Fortner, C. K., Kershaw, D. C., Purtell, K. M., & Zulli, R. A. (2010). *Portal report: Teacher preparation and student test scores in North Carolina*. North Carolina: Carolina Institute for Public Policy.

Henry, G. T., Purtell, K. M., Bastian, K. C., Fortner, C. K., Thompson, C. L., Campbell, S. L., & Patterson, K. M. (2014). The effects of teacher entry portals on student achievement. *Journal of Teacher Education*, *65*(1), 7-23. 10.1177/0022487113503871

History of Education. (2015, July 27). *Welcome to the history of education.* Retrieved from http://historyeducationinfo.com/

Holt, J. C. T. (2014). *A comparison between factor analysis and item response theory modeling in scale analysis*. Retrieved from http://www.rug.nl/research/portal/files/13080475/20140623_Gmw_TenHolt.pdf

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, *2*(1), 41-54. doi: 10.1007/s11336-011-9218-4.

Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modelling: Guidelines for determining model fit. *Articles*, *6*(1). doi:10.1016/j.paid.2006.09.018

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55. http://www.ejbrm.com/vol6/v6-i1/v6-i1-papers.htm

Hutchinson, S. R., & Olmos, A. (1998). Behavior of descriptive fit indexes in confirmatory factor analysis using ordered categorical data. *Structural Equation Modeling: A Multidisciplinary Journal*, *5*(4), 344-364. http://dx.doi.org/10.1080/10705519809540111

Ingersoll, R., Merrill, L., & May, H. (2014). *What are the effects of teacher education and preparation on beginning teacher attrition*? Retrieved from http://www.cpre.org/sites/default/files/researchreport/2018_prepeffects2014.pdf

Jaus, V. P. (1999). *Using the INTASC standards to understand and analyze the performance problems of student teachers*. Retrieved from intasc_model_core_teaching_standards_2011%20(1).pdf

Jöreskog, K. G. & Sörbom, D. (1996). *LISREL 8 User's Reference Guide*. Retrieved from http://www.ssicentral.com/lisrel/techdocs/Session1.pdf

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1-73. doi: 10.1111/jedm.12001

Kim, S., Micek, T., & Grigsby, Y. (2013). Investigating professionalism in ESOL teacher education through critical incident analysis and evaluation. *International Journal*, *2*(3), 170-186. http://dx.doi.org/10.1093/elt/ccm072

Konkolÿ Thege, B., Kovács, É., & Balog, P. (2014). A bifactor model of the Posttraumatic Growth Inventory. *Health Psychology and Behavioral Medicine: An Open Access Journal*, *2*(1), 529-540. doi: 10.1080/21642850.2014.905208

LaBue, A. C. (1960). Teacher certification in the United States: A brief history. *Journal of Teacher Education*, *11*(2), 147-172.

Lanier, J. E. (1986). *Tomorrow's teachers: A report of the Holmes Group*. Retrieved from http://files.eric.ed.gov/fulltext/ED399220.pdf

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research*, *32*(1), 53-76. http://dx.doi.org/10.1207/s15327906mbr3201_3

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology*, *51*(1), 201-226. doi: 10.1146/annurev.psych.51.1.201

MacIntosh, R., & Hashim, S. (2003). Variance estimation for converting MIMIC model parameters to IRT parameters in DIF analysis. *Applied Psychological Measurement*, *27*(5), 372-379. doi:10.1177/0146621603256021

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*, *103*(3), 391. http://dx.doi.org/10.1037/0033-2909.103.3.391

McDonald, R. P. (1989). An index of goodness-of-fit based on noncentrality. *Journal of Classification*, *6*(1), 97-103. doi: http://dx.doi.org/10.1007/BF02294359

McIntyre, D. J., & Killian, J. E. (1987). The influence of supervisory training for cooperating teachers on preservice teachers' development during early field experiences. *The Journal of Educational Research*, *80*(5), 277-282. 10.1080/00220671.1987.10885767

Merç, A. (2015). Assessing the performance in EFL teaching practicum: Student teachers' views. *International Journal of Higher Education*, *4*(2), 44. doi http://dx.doi.org/10.5430/ijhe.v4n2p44

Miles, P., & House, D. (2015). The tail wagging the dog: An overdue examination of student teaching evaluations. *International Journal of Higher Education*, *4*(2), 116. http://dx.doi.org/10.5430/ijhe.v4n2p116

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, *39*(3), 479-515. http://dx.doi.org/10.1207/S15327906MBR3903

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation, 7*(10), 71-81. doi:10.1016/j.edurev.2007.05.002

Muchinsky, P. M. (1996). The correction for attenuation. *Educational and Psychological Measurement*, *56*(1), 63-75. http://dx.doi.org/10.1037/1082-989X.10.2.206

Muthén, B. O., Kao, C. F., & Burstein, L. (1991). Instructionally sensitive psychometrics: application of a new IRT-based detection technique to mathematics achievement test items. *Journal of Educational Measurement, 28*(1), 1-22. doi:10.1177/0146621613478150

Muthén, L. K., & Muthén, B. O. (2010). *Mplus User's Guide: Statistical Analysis with Latent Variables: User's Guide*. Retrieved from https://www.statmodel.com/download/usersguide/Mplus%20Users%20Guide%20v6.pdf

Newsom, J. T. (2012). Some clarifications and recommendations on fit indices. *USP*, *655*, 123-133. oi: 10.1037/a0014694

North Carolina Professional Teaching Standards Commission. (2011)d. *North Carolina professional teaching standards.* Retrieved from http://www.ncpublicschools.org/docs/effectiveness-model/ncees/standards/prof-teach-standards.pdf

Ochieng'Ong'ondo, C., & Borg, S. (2011). “We teach plastic lessons to please them”: The influence of supervision on the practice of English language student teachers in Kenya. *Language Teaching Research*, *15*(4), 509-528. doi: 10.1177/1362168811412881

Oliden, P. E. (2011). Assessing measurement equivalence in ordered-categorical data. *Psicológica: Revista de metodología y psicología experimental*, *32*(2), 403-421. http://dx.doi.org/10.1037/a0029315

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (Vol. 161). Washington D.C.: Sage Publications.

Pearson Education. (2016, Feb 9th). *About EdTPA.* Retrieved from http://www.edtpa.com/PageView.aspx?f=GEN_AboutEdTPA.html

Pecheone, R. L., & Chung, R. R. (2006). Evidence in teacher education: The performance assessment for California teachers (PACT). *Journal of Teacher Education*, *57*(1), 22-36. doi: 10.1177/0022487105284045

Pentz, M. A., & Chou, C. P. (1994). Measurement invariance in longitudinal clinical research assuming change from development and intervention. *Journal of Consulting and Clinical Psychology*, *62*(3), 450. doi:10.1037//0022-006X.62.3.450

Ployhart, R. E., & Oswald, F. L. (2004). Applications of mean and covariance structure analysis: Integrating correlational and experimental approaches. *Organizational Research Methods*, *7*(1), 27-65. doi: 10.1177/1094428103259554

Porter, J. M., & Jelinek, D. (2011). Evaluating inter-rater reliability of a national assessment model for teacher performance. *International Journal of Educational Policies*, *5*(2), 74-87. doi:10.1177/0022429405053002071O.1177

Proitsi, P., Hamilton, G., Tsolaki, M., Lupton, M., Daniilidou, M., Hollingworth, P., ... & Todd, S. (2011). A multiple indicators multiple causes (MIMIC) model of behavioural and psychological symptoms in dementia (BPSD). *Neurobiology of Aging*, *32*(3), 434-442. doi:10.1016/j.neurobiolaging.2009.03.005

Raths, J., & Lyman, F. (2003). Summative evaluation of student teachers an enduring problem. *Journal of Teacher Education*, *54*(3), 206-216. doi:10.1002/jrsm.12

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment*, *92*(6), 544-559. doi: 10.1080/00223891.2010.496477

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*(4), 507-514. doi:10.1007/BF02296192

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika*, *75*(2), 243-248. doi:10.1007/s11336-009-9135-y

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 1-69. doi: 10.1002/j.2333-8504.1968.tb00153

Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing*, *26*(1), 005-30. http://dx.doi.org/10.1002/j.2333-8504.2008.tb0209

Schaie, K. W., Maitland, S. B., Willis, S. L., & Intrieri, R. C. (1998). Longitudinal invariance of adult psychometric ability factor structures across 7 years. *Psychology and Aging*, *13*(1), 8. doi:10.1037/0882-7974.13.1.8

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464. doi:10.2307/2958889

SelormSosu, E., Paddy, L. P., AsantewaaMintah-Adade, E., & Ativui, R. *Perceptions of history student-teachers on teaching practice supervision.* Retrieved from 14845-17490-1-PB.pdf

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4-14. doi: 10.3102/0013189X015002004

Simpson, L. J. (2004). *Student teacher exit portfolios: Is it an appropriate measure and a unique contribution toward the assessment of highly qualified teacher candidates*? Retrieved from http://drum.lib.umd.edu/bitstream/handle/1903/1377/umi-umd-1377.pdf;sequence=1

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15*(1), 72-101. doi: 10.1093/ije/dyq191

Suhr, D. (2006). *The basics of structural equation modeling*. Retrieved from http://www.lexjansen.com/wuss/2006/tutorials/TUT-Suhr.pdf

Taskstream. (2016, July 12). *About*. Retrieved from https://www1.taskstream.com/

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, 4(1), 25–29. http://dx.doi.org/10.1037/h0071663

Tobin, K. (2012). Control of teacher certification in the United States. *Peabody Journal of Education*, *87*(4), 485-499. doi:10.17226/12882

Tremblay, P. F., & Gardner, R. C. (1996). On the growth of structural equation modeling in psychological journals. *Structural Equation Modeling: A Multidisciplinary Journal*, *3*(2), 93-104. http://dx.doi.org/10.1080/10705519609540035

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*(1), 1-10. doi:10.1007/BF02291170

UNCG School of Education. *History timeline.* (2015, July 27). Retrieved from http://soe.uncg.edu/about-us/uncg-school-of-education-history-timeline/

Van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*,*9*(4), 486-492. http://dx.doi.org/10.1080/17405629.2012.68674

Voss, T., Kunter, M., & Baumert, J. (2011). Assessing teacher candidates' general pedagogical/psychological knowledge: Test construction and validation. *Journal of Educational Psychology*, *103*(4), 952. doi: 10.1037/a0025125

Walker, C. M. (2001). A review of DIFPACK: Dimensionality-based DIF analysis package. *International Journal of Testing*, *1*(3-4), 305-317. http://dx.doi.org/10.1080/15305058.2001.9669477

Wiesner, M., & Schanding, G. T. (2013). Exploratory structural equation modeling, bifactor models, and standard confirmatory factor analysis models: Application to the BASC-2 behavioral and emotional screening system teacher form. *Journal of School Psychology*, *51*(6), 751-763. http://dx.doi.org/10.1016/j.paid.2014.08.018

Westland, J. C. (2010). Lower bounds on sample size in structural equation modeling. *Electronic Commerce Research and Applications*, *9*(6), 476-487. doi:10.1016/j.elerap.2010.07.003

Wolming, S., & Wikström, C. (2010). The concept of validity in theory and practice. *Assessment in Education: Principles, Policy & Practice*, *17*(2), 117-132. http://dx.doi.org/10.1080/09695941003693856

Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research*, *44*(1), 1-27. doi: 10.1080/00273170802620121

Xie, J., Bi, Q., Shang, W., Yan, M., Yang, Y., Miao, D., & Zhang, H. (2012). Positive and negative relationship between anxiety and depression of patients in pain: A bifactor model analysis. *PloS one*, *7*(10). doi: 10.1371/journal.pone.0047577.

Yang, Y., Sun, Y., Zhang, Y., Jiang, Y., Tang, J., Zhu, X., & Miao, D. (2013). Bifactor item response theory model of acute stress response. *PloS One, 8*(6). doi:10.1371/journal.pone.0065291

Yu, C. Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes.* Retrieved from https://www.statmodel.com/download/Yudissertation.pdf

APPENDIX A

A DESCRIPTION OF THE ITEMS THAT MAKE UP EACH FACTOR IN THE UNCG ASSESSMENT SYSTEM

The TGAP

1	Long range planning with sequencing	Planning (INTASC 1, 4, 6, 7, 9)
2	alignment with curriculum	
3	material/equipment	
4	context of the lesson	Instruction (INTASC 1, 2, 3, 4, 6, 7, 8)
5	content knowledge; presentation	
6	appropriateness of lesson; pacing	
7	Use of technology and instructional resources	
8	Effectiveness of instructional resources	
9	strategies for differentiation	
10	questioning techniques	
11	Analysis of student assessment results	Assessment (INTASC 2, 5, 6, 8)
12	Meaningful student work assignments	
13	Quality of feedback to students	
14	Expectations/procedures	Student Motivation and Management (INTASC 2, 3, 5, 10)
15	Expectations for student success	
16	student interest and participation	
17	student collaboration	
18	classroom climate	

The CDAP

1	Ethical Behavior	Dispositions
2	Responsibility	
3	Personal and Professional Conduct	
4	Inclusive and affirming of diversity	
5	Collaborative	
6	Reflective Learner	
7	Receptive to Feedback	
8	Self-efficious	
9	Engaged and committed to teaching as a profession	

The Evidence Portfolios

Content	1	Identified and articulates a topic in his or her academic field	Evidence 2 In-depth Inquiry Project (content investigation, such as research paper or performance) NCPTS 3b.1
	2	Significance of topic is explained	
Depth	3	Demonstrates deep knowledge of content and complexity of topic	
Rigor	4	Literature is rich, current, cited and relevant	
	5	Collects and used data from a wide variety of sources	
	6	Depth continues to the level of understanding relationship and contradictions among interpretation surrounding the topic	
	7	Draws conclusions that reflect integration of complex data and independent critical thinking	
Presentation	8	Presentation is clear appropriate to the discipline and communicates complex ideas smoothly	
	1	Units and plans aligned with NC standard course of Study	Evidence 3 Ped. Knowledge and Skills: Planning (Unit plan, lesson plans and reflection) NCPTS 1a.2, 1a.3, 2b.1, 2b.2, 2b.3, 2d.1, 3a.1, 3a.2, 3c.1, 3c.2, 3d.1, 4a.1, 4a.2, 4b.1, 4c.1, 4d.1, 4e.1, 4f.1, 4h.1, 5c.1
	2	Formative assessment to help students progress	
	3	Summarize assessments to evaluate students	
	4	Lessons are culturally varied to help a diverse set of students	
	5	Opportunities are provided to students to expose them to and help them understand multiple points of view	
	6	The relevance of the content is explained and addressed	

	7	Multiple sources of data are used in making instructional decisions	
	8	Units are constructed to accommodate diversity special needs and students who speak English as a second language	
	9	Plans and Units integrate technology well	
	10	Plans and Units integrate multiple elements that help students apply processes for critical thinking and problem solving	
	11	Multiple activities are included to help students develop leadership ethics accountability adaptability personal productivity responsibility interpersonal skills, self-direction and social responsibility	
	12	Lesson plans foster a safe welcoming and orderly classroom	
	13	Plans reflect collaboration with colleagues and specialists	
	14	Reflections are comprehensive and logical and reflect critical thinking	
	15	The product is presented clearly, is comprehensive and is well organized	
Teaching Content	1	Overview is clear and makes connections to the specific characteristics of the school that supports making instructional decisions for all students	Evidence 5 Impact on Student Learning (positive impact on student learning) NCPTS 1a.1, 1a.2, 4a.1, 4a.2, 4e.12, 4b.1, 4h.1, 4h.2, 5a.1
	2	Assessment data related to multiple characteristics of student are analyzed and explained	

	3	Relevant subgroups as appropriately identified	
Instructional Goals and Objectives	4	Instructional goals and objectives specified in line with the NC standard course of study	
Plans for Assessment	5	Pre and post assessment with a plan for feedback to students are identified	
	6	Multiple types of assessment are used and aligned in a way the students can be actively engaged in the assessment process	
	7	Assessment measures are tailored to fit the needs of diverse students with attention to multiple characteristics of learners and appropriate adaptions	
	8	Method and timetable for collected data are described and explained	
	9	Rationale and goals for the assessment measures are provided	
Data collection and analysis (whole class)	10	Pre-assessment data are presented and the analysis of this data along with its relationship to instructional goals and plans are considered	
	11	A complete plan is present for multiple formative assessments	
	12	Summative data is used and discussed to present evidence of a positive impact on student learning.	
Data collection and analysis (subgroups)	13	Pre-assessment data are presented and the analysis of this data along with its relationship to instructional goals and plans are considered	

	14	A complete plan is present for multiple formative assessments	
	15	Summative data is used and discussed to present evidence of a positive impact on student learning.	
Instructional Monitoring and Lesson Adaptations	16	multiple lesson adaptations are explained and connected to the assessment data	
	17	instructional adaptations reflect collaboration with specialists	
	18	A summary is provided to address the strengths and weaknesses of all students including special needs	
Reflection	19	Reflections are comprehensive and logical and reflect critical thinking about the project	
Presentation	20	The product is presented clearly, is comprehensive and is well organized	
	1	Identifies the characteristics of the school improvement plan in light of the community needs	Evidence 6 Leadership Advocacy and Professional Practice (individual or group project that addresses collaboration and leadership for school improvement professional development or family involvement) NCPTS 1b.1, 1b.2, 1b.3, 1c.1, 1c.2, 2e.1, 5b.1
	2	Uses appropriate data from multiple sources to address SIP to promote student growth	
	3	Identifies the benefits for the students of the project	
	4	Uses data from multiple sources to support and evaluate the plan for the project	
	5	provides evidence of engagement with high quality professional development	
	6	Provides evidence of multiple collaboration and beginning to develop a professional network	

	7	Evidence of communication with home and communities for the benefit of the students	
	8	Organized and well developed project.	
	1	Transcript 24+ semester hours and Praxis 2 passing scores	Evidence 1 Breadth of Knowledge (Pass/Fail) NCPTS 3b.1
	1	Successful student teaching experience as outlined by the Certification of Teaching capacity	Evidence 4 Pedagogical Knowledge and Skills: Clinical Performance (Pass/Fail) NCPTS 1a.1, 1a.3, 1a.4, 1d.1, 1e.1, 2a.1, 2b.1, 2b.2, 2c.1, 2d.1, 2d.2, 3a.2, 3b.2, 3d.1, 4c.1, 4d.1, 4e.1, 4f.1, 4g.1, 4g.2, 4h.1, 4b.2, 5a.1

APPENDIX B

DESCRIPTIVE STATISTICS FOR THE ITEMS IN THE SUPERVISOR AND OSTE GROUPS

Supervisor Group descriptive statistics

N=560	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
D1	5.24	0.98	-1.13	0.10	0.14	0.21
D2	5.14	1.05	-0.93	0.10	-0.44	0.21
D3	5.19	1.01	-1.02	0.10	-0.17	0.21
D4	5.11	1.00	-0.86	0.10	-0.42	0.21
D5	5.16	1.01	-0.93	0.10	-0.36	0.21
D6	5.17	1.01	-0.99	0.10	-0.19	0.21
D7	5.25	1.01	-1.16	0.10	0.12	0.21
D8	5.08	1.01	-0.79	0.10	-0.53	0.21
D9	5.23	0.99	-1.12	0.10	0.10	0.21
T1	4.97	0.99	-0.50	0.10	-0.92	0.21
T2	5.13	0.99	-0.78	0.10	-0.59	0.21
T3	5.13	0.98	-0.76	0.10	-0.64	0.21
T4	4.90	0.97	-0.35	0.10	-1.01	0.21
T5	4.90	0.96	-0.31	0.10	-1.04	0.21
T6	4.85	0.97	-0.27	0.10	-1.05	0.21
T7	4.93	1.01	-0.37	0.10	-1.12	0.21
T8	4.93	0.99	-0.41	0.10	-1.00	0.21
T9	4.81	1.00	-0.22	0.10	-1.12	0.21
T10	4.78	0.97	-0.14	0.10	-1.09	0.21
T11	4.78	0.98	-0.17	0.10	-1.09	0.21
T12	4.99	1.00	-0.51	0.10	-0.96	0.21
T13	4.88	1.00	-0.31	0.10	-1.11	0.21
T14	5.01	0.97	-0.51	0.10	-0.88	0.21
T15	5.03	0.98	-0.57	0.10	-0.85	0.21
T16	5.01	0.98	-0.54	0.10	-0.89	0.21
T17	4.87	0.96	-0.27	0.10	-1.05	0.21
T18	5.12	0.95	-0.73	0.10	-0.58	0.21

OSTE Group Descriptive Statistics

N=560	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
D1	5.27	1.01	-1.14	0.10	-0.01	0.21
D2	5.14	1.06	-0.91	0.10	-0.55	0.21
D3	5.21	1.03	-1.03	0.10	-0.26	0.21
D4	5.10	1.04	-0.80	0.10	-0.66	0.21
D5	5.13	1.03	-0.86	0.10	-0.56	0.21
D6	5.16	1.04	-0.93	0.10	-0.47	0.21
D7	5.23	1.04	-1.06	0.10	-0.27	0.21
D8	5.02	1.04	-0.66	0.10	-0.84	0.21
D9	5.20	1.03	-1.02	0.10	-0.28	0.21
T1	5.08	0.99	-0.61	0.10	-0.91	0.21
T2	5.24	0.96	-0.95	0.10	-0.33	0.21
T3	5.17	1.00	-0.83	0.10	-0.60	0.21
T4	5.03	1.01	-0.55	0.10	-0.98	0.21
T5	4.94	1.00	-0.40	0.10	-1.09	0.21
T6	4.96	1.02	-0.45	0.10	-1.07	0.21
T7	5.16	0.98	-0.78	0.10	-0.66	0.21
T8	4.99	0.99	-0.45	0.10	-1.04	0.21
T9	4.93	1.02	-0.35	0.10	-1.18	0.21
T10	4.92	1.00	-0.36	0.10	-1.10	0.21
T11	4.86	1.01	-0.27	0.10	-1.18	0.21
T12	5.14	0.98	-0.75	0.10	-0.68	0.21
T13	5.06	0.99	-0.59	0.10	-0.91	0.21
T14	5.01	1.04	-0.56	0.10	-1.03	0.21
T15	5.20	0.98	-0.89	0.10	-0.47	0.21
T16	5.07	1.01	-0.61	0.10	-0.93	0.21
T17	4.96	1.02	-0.38	0.10	-1.19	0.21
T18	5.18	0.98	-0.81	0.10	-0.62	0.21